Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

RANKING SINGLE NUCLEOTIDE POLYMORPHISMS WITH SUPPORT VECTOR REGRESSION IN CONTINUOUS PHENOTYPES

by
Seif Shahidain

Support vector machines (SVM) have been used to improve the ranking of single nucleotide polymorphisms (SNPs) over traditional chi-square tests in disease case studies [2]. In this investigation, ranking SNPs with support vector regression (SVR) was compared to the Wald test in predicting continuous phenotypes. SVR-ranked SNPs consistently outperformed the Wald test-ranked SNPs to provide a more accurate prediction of the phenotype with fewer SNPs across several methods of prediction.

RANKING SINGLE NUCLEOTIDE POLYMORPHISMS WITH SUPPORT VECTOR REGRESSION IN CONTINUOUS PHENOTYPES

by
Seif Shahidain

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Department of Mathematical Sciences

May 2011

APPROVAL PAGE

RANKING SINGLE NUCLEOTIDE POLYMORPHISMS WITH SUPPORT VECTOR REGRESSION IN CONTINUOUS PHENOTYPES

Seif Shahidain

Dr. Usman Roshan, Thesis Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Assistant Professor of Computer Science, NJIT

Dr. Sunil Dhar, Committee Member — Date
Associate Professor of Mathematics, NJIT

BIOGRAPHICAL SKETCH

Author: Seif Shahidain

Degree: Master of Science

Date: May 2011

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2011

- Bachelor of Science in Applied Mathematics,
  Emory University, Atlanta, GA, 2009

Major: Computational Biology

Presentations and Publications:

Murali Gururajan, Trivikram Dasu, Seif Shahidain, C. Darrell Jennings, Darrell A. Robertson, Vivek M. Rangnekar and Subbarao Bondada. (2007) Spleen Tyrosine Kinase (Syk), a Novel Target of Curcumin, is Required for B Lymphoma Growth. *The Journal of Immunology*. **178** (1), 111-121.

I dedicate this to my family.

ACKNOWLEDGMENT

I would like to thank Dr. Usman Roshan for helping me, someone that did not initially know how to write scripts, perform research in the field of Bioinformatics. It truly was an illuminating experience. Also, I would like to thank Dr. Zhi Wei and Dr. Sunil Dhar for reviewing and helping me clarify my thesis work. Finally, I would like to thank my family for always supporting my decisions.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION……..…….	1
1.1 Background Information …….....……..……...	1
1.2 Objective …….…….……....	2
2 METHODS ……	3
2.1 Data …….	3
2.2 Significant SNP Selection with Wald Test …….	4
2.3 Support Vector Regression …….	4
2.4 Predicting Phenotypes …….	5
2.4.1 Ridge Regression ……..	6
2.4.2 Multiclass Support Vector Machines ……	6
3 RESULTS ……..	7
4 DISCUSSION …….....	14
4.1 Increasing the Number of SNPs Selected ……...…	14
4.2 Effective Use of Pedigree …….……..…..	15
5 CONCLUSION …….……...	16
APPENDIX A COAT COLOR ANALYSIS ……..……..	17
APPENDIX B MEAN CELLULAR HEMOGLOBIN ANALYSIS ...……..……...	23

TABLE OF CONTENTS
(Continued)

Chapter **Page**

APPENDIX C PERCENTAGE OF $CD8^+$ CELLS ANALYSIS …….. 28

APPENDIX D ANALYSIS OF OTHER PHENOTYPES …… 32

APPENDIX E RIDGE REGRESSION RESULTS ……... 43

REFERENCES ……... 44

LIST OF TABLES

Table		Page
2.1	Number of Mice in Training and Test Datasets ……...…….	3
3.1	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values using Ridge Regression with $\lambda = 5$……...…….	7
3.2	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values using Support Vector Regression ……...…….…….	8
3.3	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Clustered SNPs with Multiple Predicting Methods ...……..……	10
A.1	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Ridge Regression with no Pedigree Included at Various Values of λ ....……..	17
A.2	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Ridge Regression with Pedigree Included at Various Values of λ ……..…….	17
A.3	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Support Vector Regression with Pedigree Included at Various Values of C …...……..…….	19
A.4	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Multiclass SVM with Pedigree Included at Various Values of C.....…….……..	19
B.1	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Ridge Regression with no Pedigree Included at Various Values of λ ...……..……	23
B.2	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Ridge Regression with Pedigree Included at Various Values of λ …..……...…….	23
B.3	Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Support Vector Regression with Pedigree Included at Various Values of C ……..……...	26

LIST OF TABLES
(CONTINUED)

Table **Page**

C.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Ridge Regression with no Pedigree Included at Various Values of λ 28

C.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Ridge Regression with Pedigree Included at Various Values of λ 28

C.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Support Vector Regression with Pedigree Included at Various Values of C 31

D.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Biochemical Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$ 32

D.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Immunological Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$ 36

D.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Hematological Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$ 39

E.1 Actual and Predicted Values of Phenotypes with Ridge Regression 43

LIST OF FIGURES

Figure		Page
3.2	Prediction of coat color, MCH and %CD8 using ridge regression with $\lambda = 5$ …..	9
3.2	Prediction of coat color, MCH and %CD8 using SVR with various values of *C*...	11
3.3	Prediction of coat color by multiclass SVM with $C = 5000$…..……...……..	12
3.4	Effect of Increasing SNPs used in SVR ranking method ……..	13
A.1	The prediction of coat color using ridge regression without including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……	18
A.2	The prediction of coat color using ridge regression including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……..……	20
A.3	The prediction of coat color using multiclass SVM including pedigree with $C = \{10, 1, .1, .01, .001, .00001\}$ ……...……...	21
A.4	The prediction of coat color using multiclass SVM including pedigree with $\lambda = \{5000, 1000, 500, 100\}$…...…….……..……...……...	22
B.1	The prediction of mean cellular hemoglobin using ridge regression without including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……	24
B.2	The prediction of mean cellular hemoglobin using ridge regression including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……	25
B.3	The prediction of mean cellular hemoglobin using support vector regression including pedigree with $C = \{0, .1, .01, .001\}$ …….	26
B.4	The prediction of mean cellular hemoglobin using support vector regression including pedigree with $C = \{.0001, .00001\}$ …….	27
C.1	The prediction of percentage of $CD8^+$ cells using ridge regression without including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……	29
C.2	The prediction of percentage of $CD8^+$ cells using ridge regression including pedigree with $\lambda = \{0, .1, 1, 5, 10, 100\}$ ……..……	30
C.3	The prediction of percentage of $CD8^+$ cells using support vector regression including pedigree with $C = \{0, .01, .001, .0001\}$ …….	31

LIST OF FIGURES
(Continued)

Figure **Page**

D.1 The prediction of Albumin, ALP, ALT, AST, Calcium, and Chloride using ridge regression including pedigree with $\lambda = 5$…….……..... 33

D.2 The prediction of Creatinine, Glucose, HDL, LDL, Phosphorus and Potassium using ridge regression including pedigree with $\lambda = 5$…….…… 34

D.3 The prediction of Sodium, Tot.Cholesterol, Tot.Protein, Triglycerides and Urea using ridge regression including pedigree with $\lambda = 5$…….…… 35

D.4 The prediction of B220Median and CD4XGeoMean using ridge regression including pedigree with $\lambda = 5$…….……..... 36

D.5 The prediction of CD4YGeoMean, CD4inCD3XGeoMean, CD4inCD3YGeoMean, CD8XGeoMean, CD8YGeoMean and PctB220 using ridge regression including pedigree with $\lambda = 5$…….…… 37

D.6 The prediction of PctCD3, PctCD4, PctCD4inCD3 and PctCD8inCD3 using ridge regression including pedigree with $\lambda = 5$…….……..…… 38

D.7 The prediction of ALYabs and BASabs using ridge regression including pedigree with $\lambda = 5$…….…….…… 39

D.8 The prediction of HCT, HGB, LICabs, LYMabs, MCHC and MCV using ridge regression including pedigree with $\lambda = 5$…….……... 40

D.9 The prediction of MONabs, MPV, NEUabs, PCT, PLT and RBC using ridge regression including pedigree with $\lambda = 5$…….……...… 41

D.10 The prediction of RDW and WBC using ridge regression including pedigree with $\lambda = 5$……...…….…… 42

CHAPTER 1

INTRODUCTION

1.1 Background

Genome-wide association studies provide insight into how specific regions of the genome affect certain diseases and phenotypes by investigating the differences between individuals with certain traits at a genetic level [1-5]. These differences between genomes are classified as single nucleotide polymorphisms (SNPs). Ranking a SNP's effect on the disease is a crucial procedure because it, not only, illuminates genes that contribute to expression of a phenotype, but can also predict certain characteristics someone will have based on their genome. The significance of a SNP is usually determined by a chi-square test in most cases. However, when dealing with a continuous phenotype, like hemoglobin level, the chi-square test is not as useful and usually the likelihood ratio test or Wald test is used to find significant SNPs [1,4].

Finding new methods that improve upon the chi-square and Wald tests are crucial to enhancing risk prediction and illuminating the regions that cause a certain trait to be expressed. Previous studies showed an improvement in disease prediction by selecting significant SNPs with support vector machine (SVM) and random forest methods. The SVM and random forest methods showed an improvement in the ranking of causal variants and associated regions over the chi-square test. This improvement, not only, enhanced the accuracy of disease risk prediction, but also reduced the number of SNPs necessary for the observed increase in predictive power [2].

1.2 Objective

The SVM and random forest results were for disease prediction, which requires the prediction of two classes: disease and no disease [2]. However, these classifier methods cannot be used with continuous phenotypes. So, a similar method of selecting significant SNPs with support vector regression (SVR) was used on data from the Wellcome Trust Centre for Human Genetics [6]. Three traits from this dataset were previously analyzed with a Bayesian method to predict phenotypes by estimating additive and dominant effects of the genotype [3]. These results were compared to the SVR method of selecting significant SNPs and predicting phenotypic values with ridge regression, SVR and multiclass SVM. In addition to these three traits, several other phenotypes were also analyzed to compare the Wald test selection of significant SNPs to the SVR method.

CHAPTER 2

METHODS

2.1 Data

The data was made publically available by the Wellcome Trust Centre for Human Genetics and contains the genotypic and phenotypic data of over 2000 mice. The data includes information on 84 families, with eight large, complex families that included a majority of the mice and 76 nuclear families [6]. The pedigree information was included as extra variables within the data by including a variable for the mouse's family and one for its parents. The mice were then randomly divided in half into a training (estimation) and test (prediction) dataset based upon their family, with at least one family member going into the training dataset. The average training and test dataset size for ten trials is included in Table 2.1. The three traits that were analyzed were coat color, percentage of $CD8^+$ cells (%CD8) and mean cellular hemoglobin (MCH), with the full results included in Appendix A, B and C respectively, while the other phenotypes analyzed are included in Appendix D.

Table 2.1 Number of Mice in Training and Test Datasets

Trait	**Total Number of Mice**	**Training Set**	**Test Set**
Coat Color	1893	965 (5)	928 (5)
MCH	1591	815 (4)	776 (5)
%CD8	1521	775 (5)	746 (5)

2.2 Significant SNP Selection with Wald Test

To test for the significance of certain SNPs, PLINK's implementation of the Wald statistic was used since it is asymptotically similar to the likelihood ratio test, the preferred method of finding significant SNPs [1,4]. The SNPs with *P*-values smaller than the Bonferroni correction; which is .05 divided by the number of SNPs, were then extracted to be used with SVR ranking and for prediction [2].

Clustering Significant SNPs

The most significant SNPs from the Wald test were then clustered based on location in the genome via k-means clustering. The k-means objective function finds the clusters, C_i, such that the following equation is minimized:

$$\sum_{i=1}^{n} \sum_{x_j \in C_i} \|x_j - m_i\|^2$$

where n is the number of clusters, x_j is the location of the SNP in the genome and m_i is the mean value of the cluster, C_i. The clusters of size 5, 10 and 20 were created and following clustering, the most significant SNP in the cluster was extracted for analysis.

2.3 Support Vector Regression

Support Vector Regression seeks to find a function, $f(x)$, that minimizes the deviation, ε, between labels, y_i, of the n training samples, given by $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset X \times \mathbb{R}$ where $X = \mathbb{R}^d$ and x_i is the SNP genotype of the i-th data point, within a certain degree of accuracy while trying to remain as flat as possible. For the linear case,

$$f = \langle w, x \rangle + b \text{ with } w \in X, b \in \mathbb{R}$$

where $\langle .,. \rangle$ denotes the dot product in X and flatness is given by a small value of w; usually established by minimizing the length of w, $\|w\| = \langle w, w \rangle$. So, the formulation of SVR becomes:

$$minimize \ \|w\| + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$subject\ to \begin{cases} y_i - f(x) \leq \varepsilon + \xi_i \\ f(x) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where $C > 0$ is the tradeoff between the tolerance to deviation and the flatness of f and $|\xi|_\varepsilon := \begin{cases} 0 & if\ |\xi| < \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases}$, known as the ε-insensitive loss function. Lagrange multipliers can be applied to the dual formulation to find w and b [7]. For more information on solving and implementing SVR refer to [7,8].

SVR-Ranked SNPs

The absolute value of the elements in the w-vector, the discriminant from SVR, is used to obtain the ranking of SNPs by sorting the entries of the w-vector in descending order [2]. The SNPs are then reordered according to the maximum absolute value entries in the SVR discriminant.

2.4 Predicting Phenotypes

In addition to Support Vector Regression, Ridge Regression and Multiclass Support Vector Machines were used to predict phenotypes from the top ranked SNPs from the Wald Test, clustering results and SVR discriminant. The accuracy of prediction was

based upon the correlation between the actual phenotypic value and the value from regression/classification.

2.4.1 Ridge Regression

Linear regression has been shown to have stability issues when the matrix (X'X) is singular. To circumvent these issues; which are caused by the high correlation between SNPS, ridge regression is used to adjust to potential linkage disequilibria. Under ridge regression the coefficient vector, β, is:

$$\beta = (X'X + \lambda I)^{-1}\text{X'Y}$$

where X is the SNP training data, Y is the phenotypic value, I is the identity matrix and λ is the ridge parameter that reduces the effect of highly correlated SNPs [5].

2.4.2 Multiclass Support Vector Machines

For multiclass phenotypes, multiclass support vector machines were used to predict phenotypes. The following optimization problem is solved for k classes:

$$minimize\ \frac{1}{2}\sum_{i=1}^{k}\|w_i\| + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$such\ that\ for\ all\ y\ in\ [1,k]\ \begin{cases} \langle x_1, w_{yi}\rangle \geq \langle x_1, w_y\rangle + 100 * \Delta(y_1, y) - \xi_1 \\ \cdots \\ \langle x_n, w_{yn}\rangle \geq \langle x_n, w_y\rangle + 100 * \Delta(y_n, y) - \xi_n \end{cases}$$

where $\Delta(y_i, y) := \begin{cases} 1 & if\ y_i = y \\ 0 & otherwise \end{cases}$ is the loss function [7]. For more information on solving the Lagrangian of the optimization problem and implementation of multiclass SVM refer to [7, 9].

CHAPTER 3

RESULTS

In the study that describes the SVM method of ranking SNPs, the results showcased that the SVM-ranked SNPs consistently outperformed the chi-square ranked SNPs by obtaining a more accurate prediction with fewer SNPs [2]. The following results are the maximum correlations obtained using the pedigrees of the mice and various values of λ and C, while further analysis of including and omitting family data as well as other values of λ and C are available in Appendices A-C. These results are compared to those in a previous study that obtained a maximum correlation between actual and predicted phenotypes of .87, .36 and .58 for coat color, mean cellular hemoglobin and percentage of $CD8^+$ cells, respectively, by using the *Reversible Jump Markov Chain Monte Carlo* (RJMCMC) to obtain various estimates for values included in the additive and dominance genetic model (Model AD) [3].

The results of ridge regression with $\lambda = 5$ are shown in Table 3.1 and Figure 3.1. The values of phenotype prediction for one trial with 850 SNPs used are shown in Appendix E since prediction deteriorated as the number of SNPs increased to 800.

Table 3.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values using Ridge Regression with $\lambda = 5$

Trait	Wald Test	Number of SNPs	SVR	Number of SNPs
Coat Color	0.64 (0.02)	235	0.64 (0.02)	110
MCH	0.39 (0.02)	250	0.39 (0.02)	85
%CD8	0.46 (0.04)	230	0.64 (0.02)	135

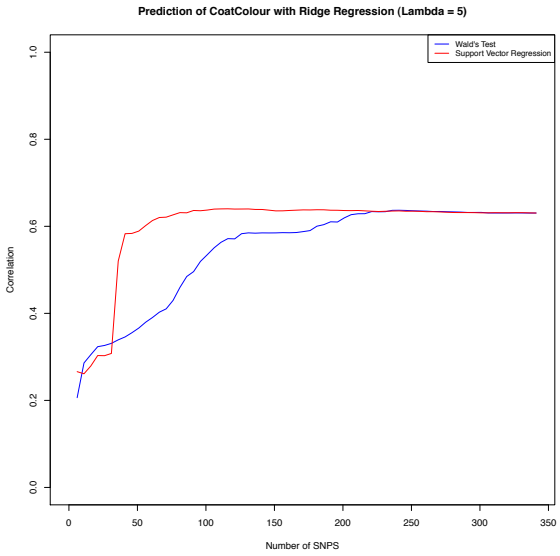
When using SVR for prediction sometimes Wald-test ranked SNPs obtained a maximum before the SVR-ranked SNPs, as shown in Table 3.2 with the values of tradeoff (*C*) that result in the highest correlation, for all values of *C* used consult Appendices A-C. However, Figure 3.2 used SVR with the same *C* values shown in Table 3.2 for prediction and confirmed that SVR-ranked SNPs produce higher correlations with fewer SNPs.

Table 3.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values using Support Vector Regression

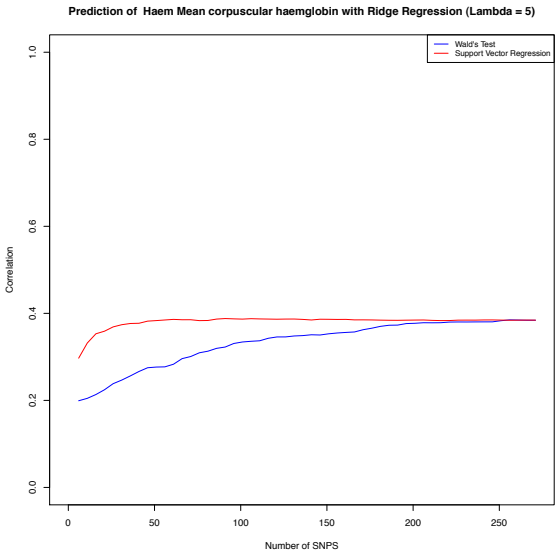
Trait	Tradeoff (*C*)	Wald Test	Number of SNPs	SVR	Number of SNPs
Coat Color	.0001	0.48 (0.02)	295	0.48 (0.02)	330
MCH	.1	0.39 (0.02)	250	0.39 (0.02)	115
%CD8	.01	0.64 (0.02)	2000	0.64 (0.01)	955

Multiclass SVM was also used for coat color prediction because coat color is a discrete phenotype, so it can be separated into distinct classes, unlike MCH and %CD8; which are continuous phenotypes. When using multiclass SVM, a similar result to SVR prediction was observed in coat color prediction using multiclass SVM with C = 5000, the Wald test-ranked SNPs and SVR-ranked SNPs both attained a maximum value of .90 (.02) with 235 and 225 SNPs, respectively, while Figure 3.3 confirms previous results of SVR-ranked SNPs having higher correlations with fewer SNPs.

A)



B)



C)

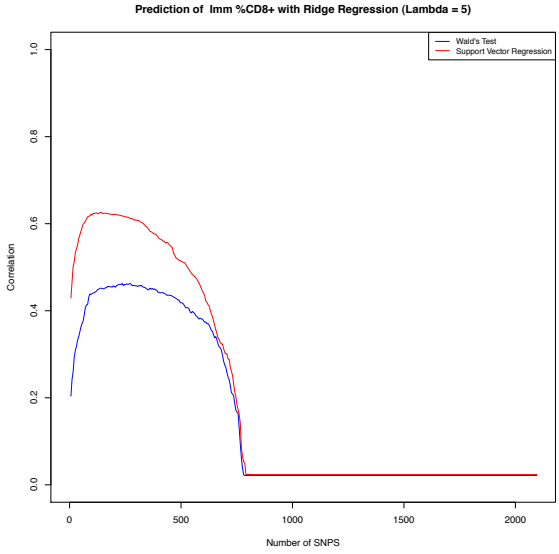


Figure 3.1 The prediction of coat color (A), MCH (B) and %CD8 (C) using ridge regression with $\lambda = 5$ shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

Table 3.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Clustered SNPs with Multiple Predicting Methods

Trait	Predicting Method	λ or C	5 Clusters	10 Clusters	20 Clusters
Coat Color	RR	5	0.39 (0.11)	0.44 (0.08)	0.44 (0.08)
	SVR	.0001	0.30 (0.07)	0.33 (0.07)	0.35 (0.06)
	Multi-SVM	5000	0.31 (0.19)	0.37 (0.16)	0.37 (0.16)
MCH	RR	5	0.26 (0.03)	0.26 (0.02)	0.27 (0.03)
	SVR	.1	0.25 (0.03)	0.26 (0.02)	0.25 (0.03)
%CD8	RR	5	0.29 (0.04)	0.33 (0.05)	0.41 (0.04)
	SVR	.01	0.26 (0.04)	0.32 (0.04)	0.40 (0.04)

The accuracy of predicted phenotypes based upon the clustering of SNPs tended to be less accurate than those obtained through the Wald test and SVR-ranked SNPs. The correlations of ridge regression, SVR and multiclass SVM using the top Wald test-ranked SNPs within 5, 10 and 20 clusters are presented in Table 3.3. Clustering did not show significant improvement over Wald test and SVR-ranked SNPs regardless of predictive method used.

The SVM method decreased progressively in ranking SNPs as the number of SNPs taken increased from r to $2r$ to $5r$ to the entire SNP genotype, where r is the number of SNPs within the Bonferroni correction, as compared to the chi-square test [2]. Figure 3.4 shows that increasing the number of SNPs taken actually improved the ranking of SNPs over the Wald test for all thresholds, however, the improvements according to certain thresholds differed across phenotypes.

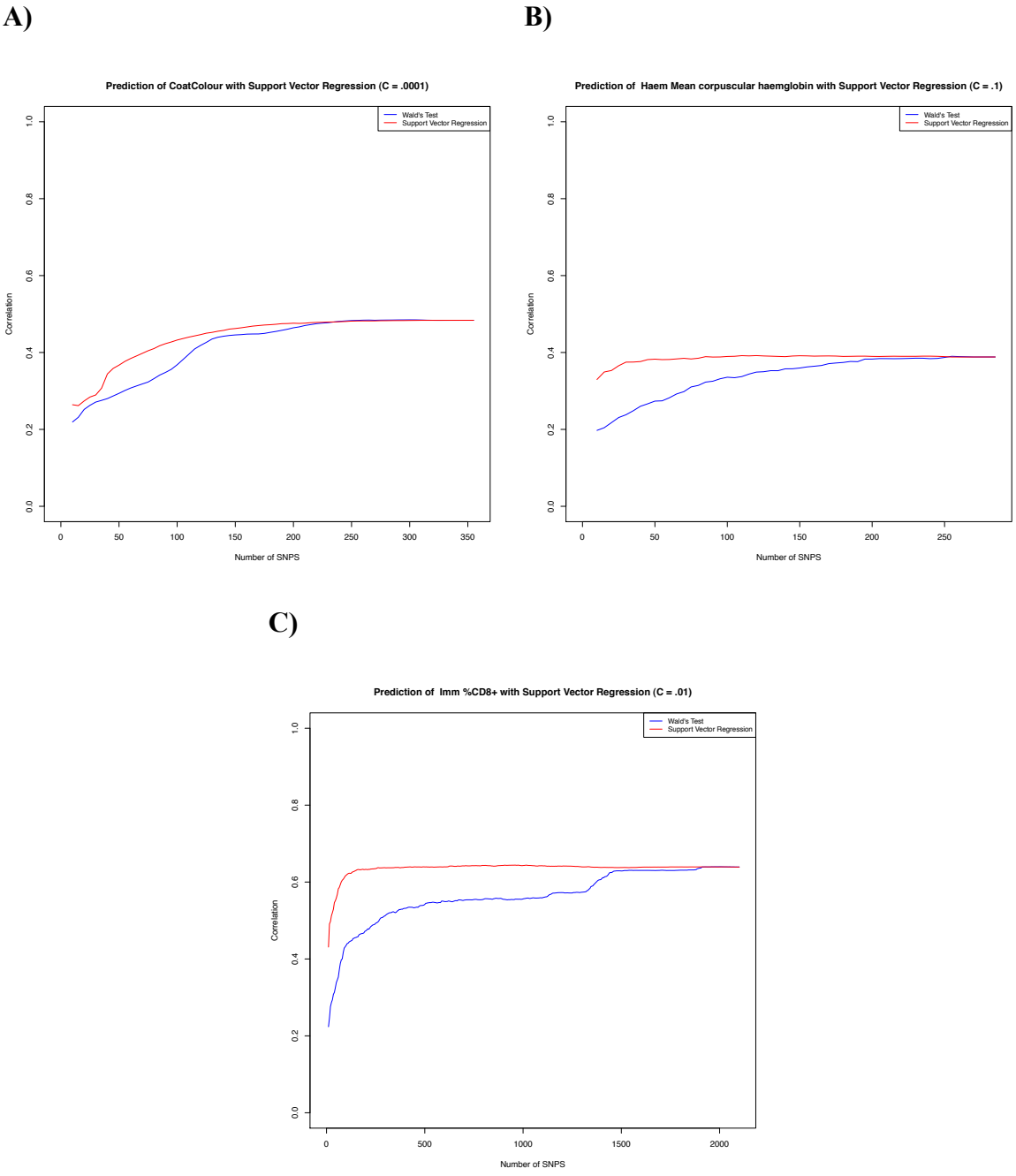


Figure 3.2 The prediction of coat color (A) with C = .0001, MCH (B) with C = .1 and %CD8 (C) with C = .01 using SVR shows SVR-ranked SNPs (red) outperforming Wald test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

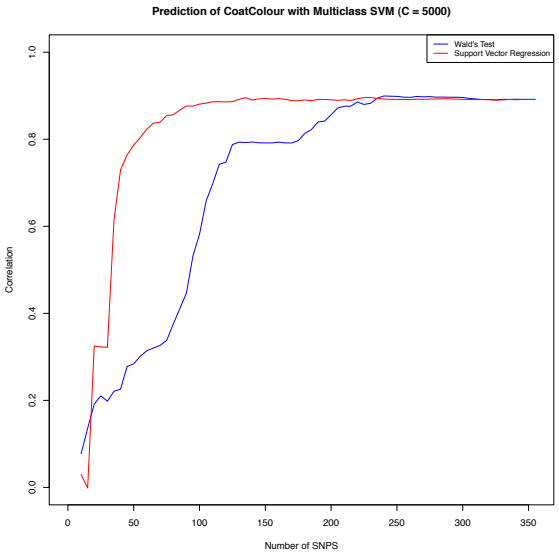
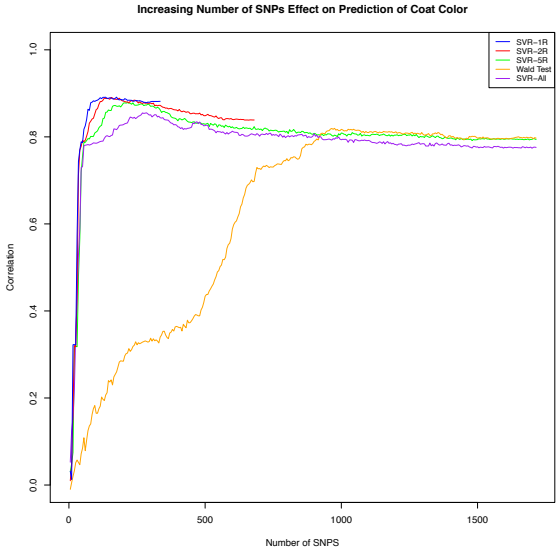


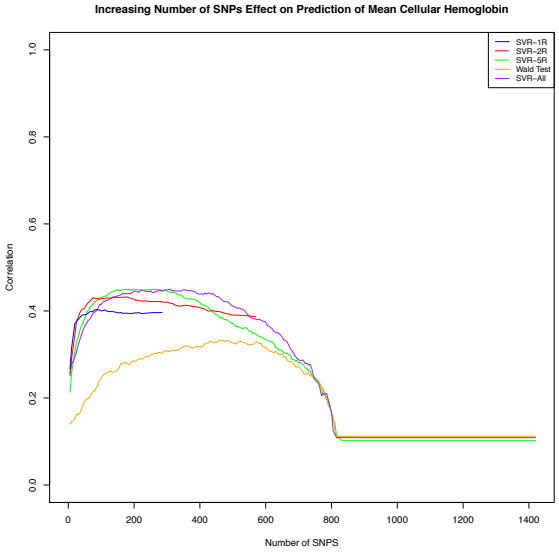
Figure 3.3 Prediction of coat color by multiclass SVM with $C = 5000$ shows that SVR-ranked SNPs (red) outperform Wald test-ranked SNPs (blue) by attaining higher correlations with fewer SNPs.

In summary, the results show that the SVR method is able to improve upon the Wald test ranking of SNPs and the results of clustering by achieving higher correlations with fewer SNPs. These results are confirmed at various values of λ for ridge regression and C for SVR and multiclass SVM; which are included in Appendices A-C, while similar results are shown for various other phenotypes in Appendix D. The best results obtained by coat color, mean cellular hemoglobin and percentage of $CD8^+$ cells; which were .90, .39 and .64, respectively, also coincided with and improved upon the results from Model AD; which attained values of .87, .36 and .58, respectively [3].

A)



B)



C)

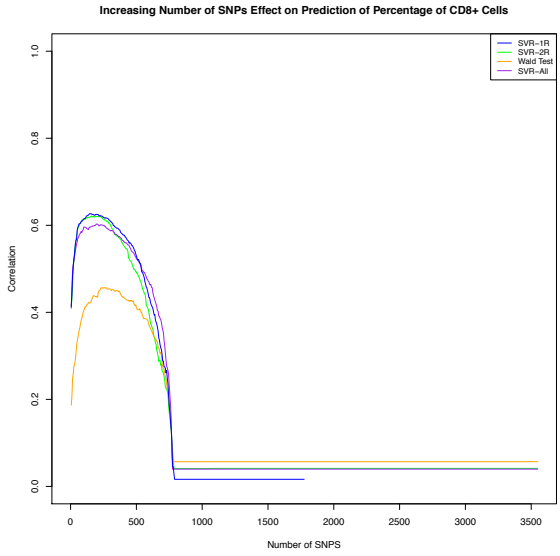


Figure 3.4 The effect of increasing the number of SNPs used in the SVR ranking method with r (blue), $2r$ (red), $5r$ (green), Wald Test (orange) and SVR method on all SNPs (purple). Prediction of Coat Color (A) with multiclass SVM (C = 5000) and MCH (B) and %CD8 (C) with ridge regression ($\lambda = 5$) showed that the SVR method improves over Wald test across all thresholds.

CHAPTER 4

DISCUSSION

The goal of this study was to examine whether the SVR-ranking method is applicable to continuous or multiple phenotypes along with a comparison to the Wald Test and clustered ranking of SNPs. The SVR-ranking method consistently achieved higher correlations with fewer SNPs as compared to the significant SNPs from the Wald test.

4.1 Increasing the Number of SNPs Selected

This study analyzed *r* SNPs in each trial, where *r* is the number of SNPs with *P*-values, attained from the Wald test, that are within the Bonferroni correction. As *r* is increased to 2*r* there was an improvement in the SVM and random forest methods detection of Type 1 diabetes-associated regions that deteriorated as the number of SNPs increased to 5*r* and 10*r* [2]. When increasing the number of SNPs selected, MCH prediction improved while a slight deterioration was observed in the Coat Color and %CD8 phenotypes. However, the SVR method consistently outperformed the Wald test at all thresholds; which was not observed previously when comparing the SVM method to the chi-square test [2]. This improvement is likely to be due to the size of the datasets used, the mouse dataset contained over 12000 SNPs while the human dataset contains over 500000 SNPs.

4.2 Effective Use of Pedigree

Included in the prediction of phenotypes were the two pedigree variables, family and parent number. With around 84 families and almost double that of parents, a problem arose in prediction when several independent family and parent variables in the mouse dataset were regressed upon a few dependent variables and vice versa. So, if the 0, 10, 30 and 80th family all had a white coat color given a value of 0 and the 5, 6, 40 and 50th families all had a black coat color given a value of 9, these phenotypes cannot be accurately portrayed with a linear model. Accordingly, a similar problem arises when a given family has a diverse phenotypic makeup. To account for the effects of pedigree, a best linear unbiased prediction was used followed by a remodeling of SNPs with every round of RJMCMC calculation [3].

CHAPTER 5

CONCLUSION

An improvement in ranking SNPs with support vector regression was observed compared to that of the selection of significant SNPs by the Wald test with SVR-ranked SNPs consistently achieving higher accuracy of phenotype prediction with fewer SNPs. This improvement was seen across all methods of phenotype prediction and the maximum correlations observed were higher than those in previous studies.

APPENDIX A

COAT COLOR ANALYSIS

This appendix contains information of the various parameters used in each of the prediction methods for coat color with the figures corresponding to the preceding tables.

Table A.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Ridge Regression with no Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.52 (.07)	200	0.62 (.02)	85	0.38 (.11)	0.44 (.08)	0.45 (.08)
.1	0.31 (.09)	105	0.43 (.07)	55	0.39 (.11)	0.44 (.08)	0.45 (.08)
1	0.58 (.03)	230	0.62 (.02)	85	0.39 (.11)	0.44 (.08)	0.44 (.08)
5	0.63 (.02)	230	0.64 (.02)	110	0.38 (.11)	0.44 (.08)	0.44 (.08)
10	0.64 (.02)	235	0.64 (.02)	100	0.39 (.11)	0.44 (.08)	0.44 (.08)
100	0.56 (.03)	240	0.59 (.02)	105	0.39 (.11)	0.44 (.08)	0.44 (.08)

Table A.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Ridge Regression with Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.53 (.07)	200	0.62 (.01)	85	0.38 (.11)	0.44 (.08)	0.44 (.08)
.1	0.31 (.09)	105	0.43 (.07)	55	0.38 (.11)	0.44 (.08)	0.44 (.07)
1	0.58 (.03)	230	0.62 (.01)	85	0.39 (.11)	0.44 (.08)	0.44 (.08)
5	0.64 (.02)	235	0.64 (.02)	110	0.39 (.11)	0.44 (.08)	0.44 (.08)
10	0.64 (.02)	235	0.64 (.01)	100	0.39 (.11)	0.44 (.08)	0.44 (.08)
100	0.53 (.02)	240	0.57 (.02)	95	0.40 (.11)	0.44 (.08)	0.44 (.08)

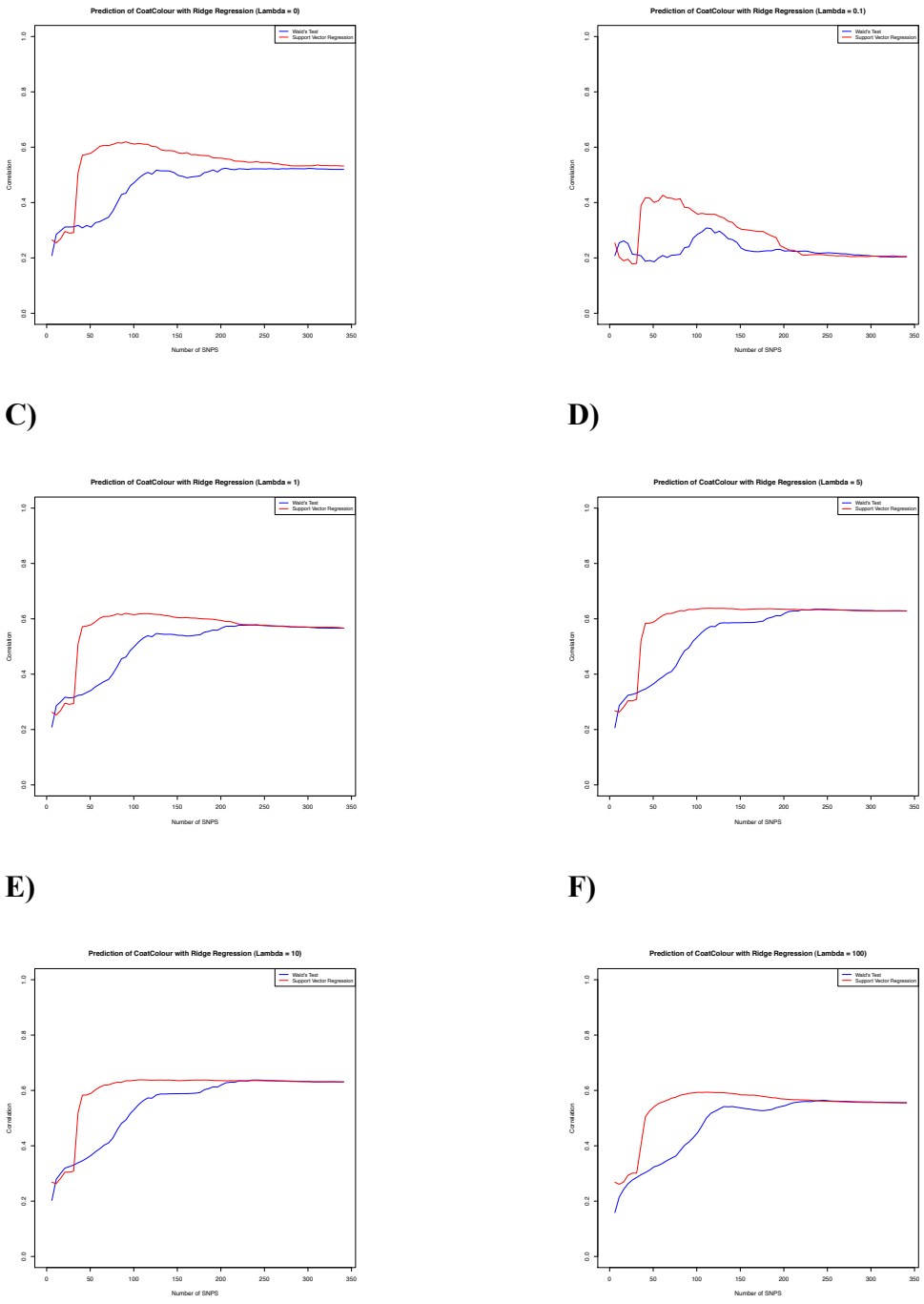


Figure A.1 The prediction of coat color using ridge regression without including pedigree with $\lambda = 0$ (A), $\lambda = 0.1$ (B), $\lambda = 1$ (C), $\lambda = 5$ (D), $\lambda = 10$ (E) and $\lambda = 100$ (F) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue).

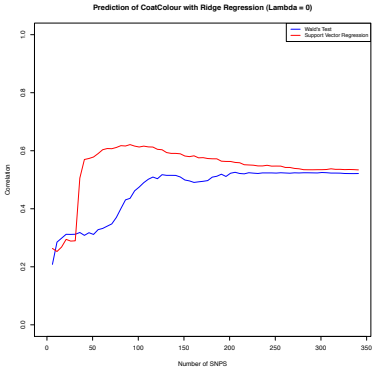
Table A.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Support Vector Regression with Pedigree Included at Various Values of *C*

C	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.53 (.07)	200	0.62 (.01)	85	0.38 (.11)	0.44 (.08)	0.44 (.08)
.1	0.31 (.09)	105	0.43 (.07)	55	0.38 (.11)	0.44 (.08)	0.44 (.07)
.01	0.58 (.03)	230	0.62 (.01)	85	0.39 (.11)	0.44 (.08)	0.44 (.08)
.001	0.64 (.02)	235	0.64 (.02)	110	0.39 (.11)	0.44 (.08)	0.44 (.08)
.0001	0.64 (.02)	235	0.64 (.01)	100	0.39 (.11)	0.44 (.08)	0.44 (.08)
.00001	0.53 (.02)	240	0.57 (.02)	95	0.40 (.11)	0.44 (.08)	0.44 (.08)

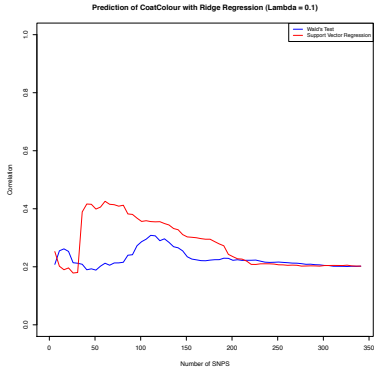
Table A.4 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Coat Color using Multiclass SVM with Pedigree Included at Various Values of *C*

C	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
5000	0.90 (.02)	235	0.90 (.02)	225	0.31 (.19)	0.37 (.16)	0.37 (.16)
1000	0.86 (.03)	245	0.86 (.03)	340	0.28 (.19)	0.34 (.14)	0.37 (.12)
500	0.83 (.02)	290	0.82 (.03)	340	0.30 (.18)	0.32 (.16)	0.36 (.12)
100	0.72 (.02)	245	0.72 (.03)	275	0.30 (.12)	0.29 (.12)	0.31 (.10)
10	0.59 (.08)	275	0.58 (.07)	290	0.29 (.15)	0.30 (.18)	0.34 (.17)
1	0.59 (.06)	165	0.59 (.07)	225	0.11 (.13)	0.25 (.13)	0.23 (.14)
.1	0.50 (.07)	245	0.50 (.10)	205	0.04 (.12)	0.09 (.11)	0.09 (.09)
.01	0.32 (.08)	305	0.31 (.06)	335	0.04 (.09)	-0.00 (.08)	0.03 (.09)
.001	0.11 (.04)	150	0.12 (.05)	230	0.05 (.09)	0.04 (.10)	0.07 (.06)
.00001	0.00 (.01)	180	0.00 (.01)	170	-0.01 (.04)	-0.01 (.04)	-0.01 (.05)

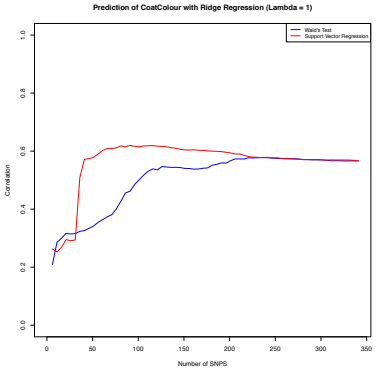
A)



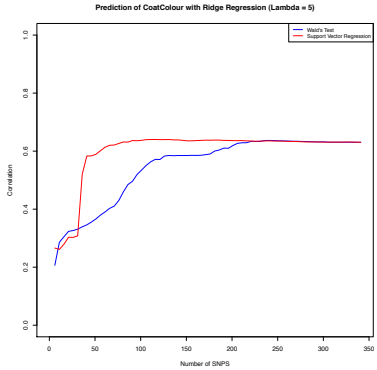
B)



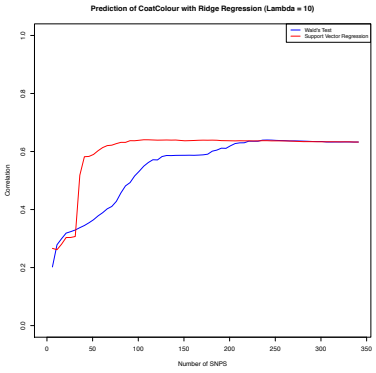
C)



D)



E)



F)

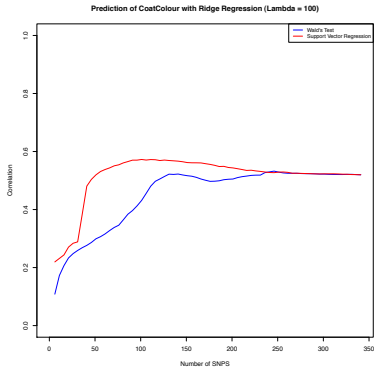
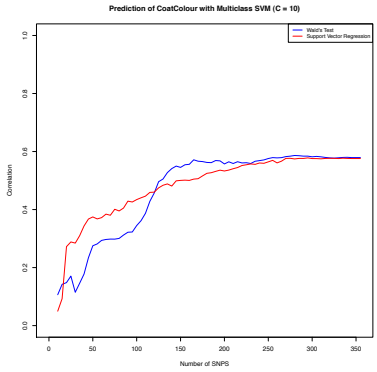
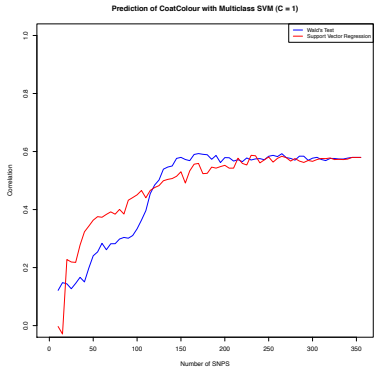


Figure A.2 The prediction of coat color using ridge regression including pedigree information with $\lambda = 0$ (A), $\lambda = 0.1$ (B), $\lambda = 1$ (C), $\lambda = 5$ (D), $\lambda = 10$ (E) and $\lambda = 100$ (F) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

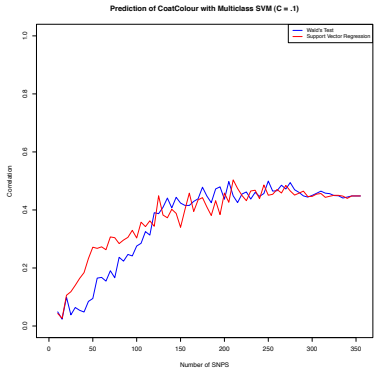
A)



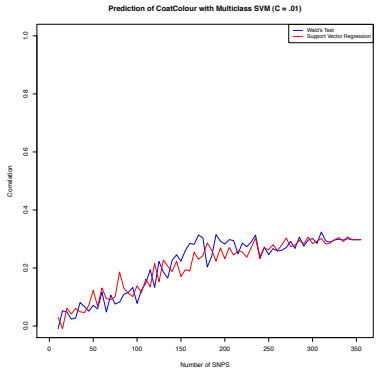
B)



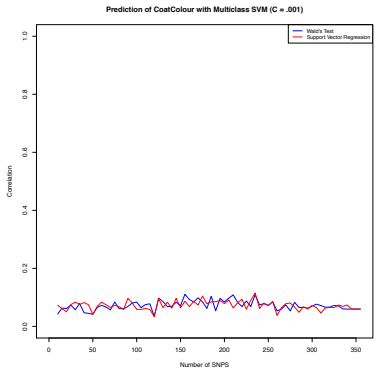
C)



D)



E)



F)

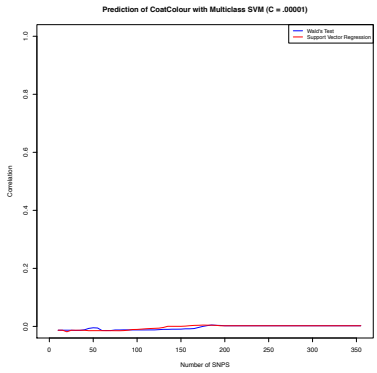


Figure A.3 The prediction of coat color using multiclass SVM including pedigree information with C = 10 (A), C = 1 (B), C = .1 (C), C = .01 (D), C = .001 (E) and C = .00001 (F) shows the improvement in prediction by increasing the trade-off, C.

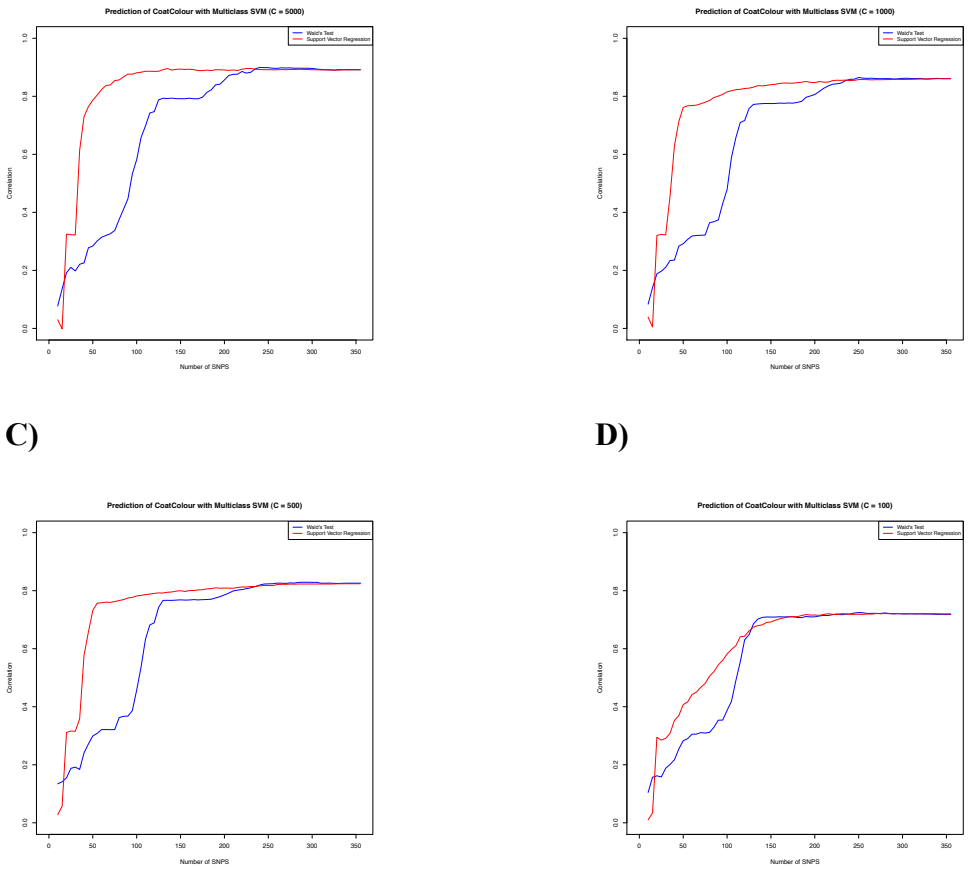


Figure A.4 The prediction of coat color using multiclass SVM including pedigree information with $C = 5000$ (A), $C = 1000$ (B), $C = 500$ (C) and $C = 100$ (D) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

APPENDIX B

MEAN CELLULAR HEMOGLOBIN ANALYSIS

This appendix contains information of the various parameters used in each of the prediction methods for mean cellular hemoglobin with the figures corresponding to the preceding tables.

Table B.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Ridge Regression with no Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.26 (.07)	115	0.36 (.02)	40	0.21 (.04)	0.23 (.03)	0.23 (.03)
.1	0.10 (.06)	0	0.29 (.06)	5	0.22 (.03)	0.23 (.03)	0.23 (.03)
1	0.31 (.03)	265	0.36 (.02)	30	0.22 (.04)	0.23 (.03)	0.23 (.03)
5	0.36 (.02)	250	0.36 (.02)	55	0.22 (.04)	0.23 (.03)	0.23 (.03)
10	0.35 (.02)	250	0.36 (.02)	55	0.22 (.03)	0.23 (.03)	0.22 (.03)
100	0.28 (.04)	260	0.29 (.03)	80	0.22 (.04)	0.23 (.03)	0.23 (.03)

Table B.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Ridge Regression with Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.29 (.06)	115	0.38 (.02)	40	0.25 (.03)	0.26 (.02)	0.26 (.02)
.1	0.14 (.07)	0	0.31 (.07)	10	0.26 (.03)	0.26 (.03)	0.27 (.03)
1	0.33 (.03)	265	0.38 (.02)	30	0.26 (.03)	0.26 (.03)	0.26 (.03)
5	0.39 (.02)	250	0.39 (.02)	85	0.26 (.03)	0.26 (.02)	0.27 (.03)
10	0.38 (.02)	250	0.38 (.02)	105	0.26 (.03)	0.27 (.03)	0.26 (.03)
100	0.30 (.02)	260	0.30 (.01)	115	0.26 (.03)	0.26 (.02)	0.26 (.03)

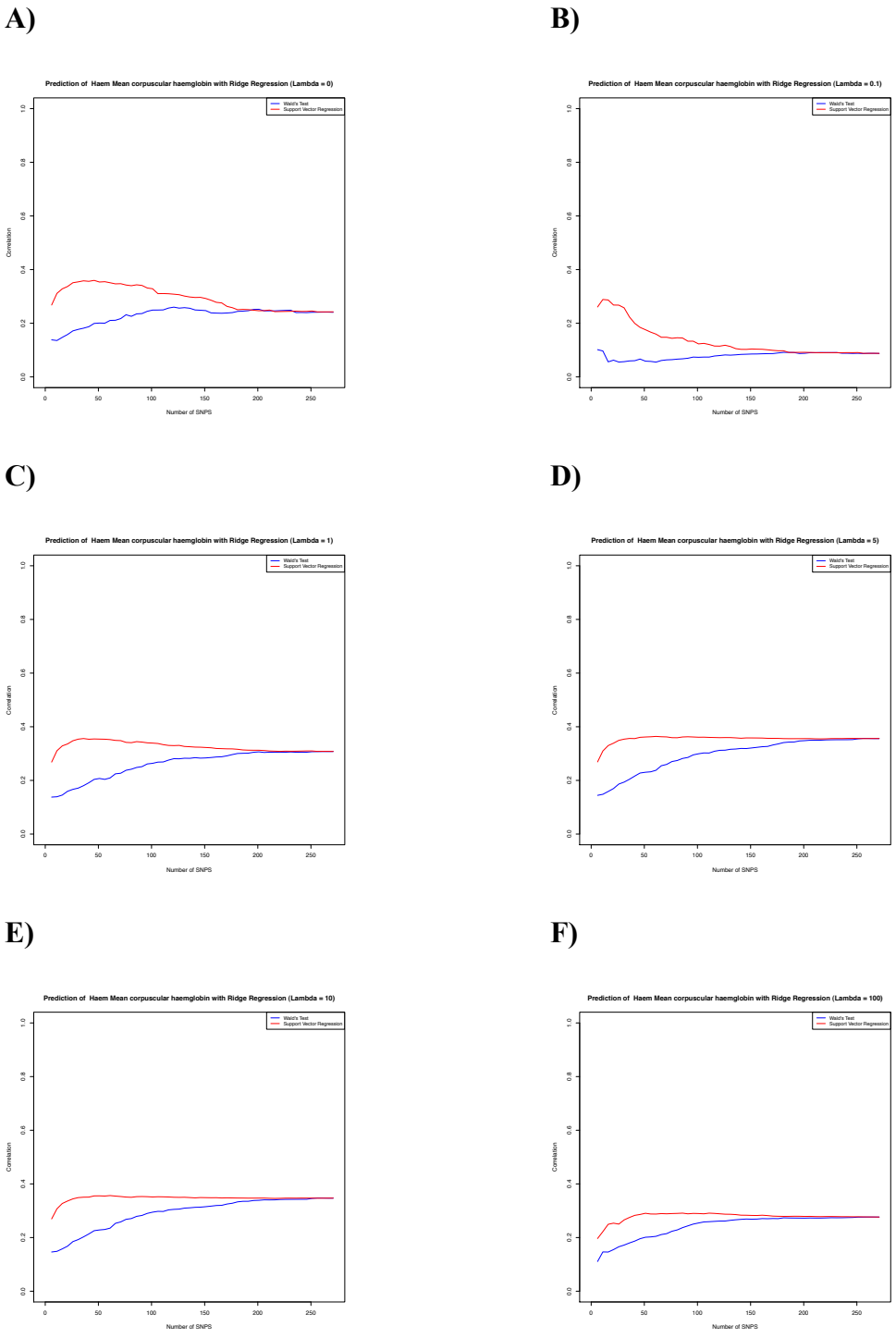
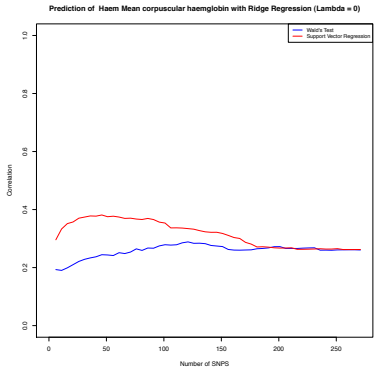
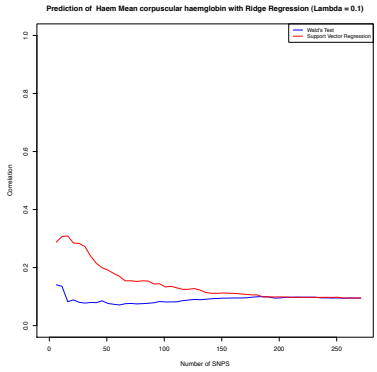


Figure B.1 The prediction of mean cellular hemoglobin using ridge regression without including pedigree with $\lambda = 0$ (A), $\lambda = 0.1$ (B), $\lambda = 1$ (C), $\lambda = 5$ (D), $\lambda = 10$ (E) and $\lambda = 100$ (F) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

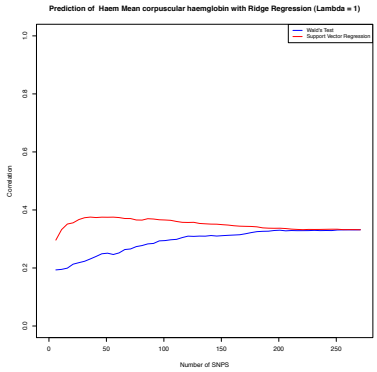
A)



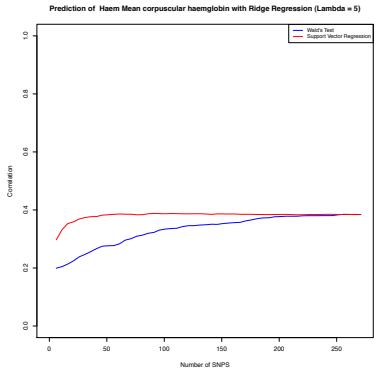
B)



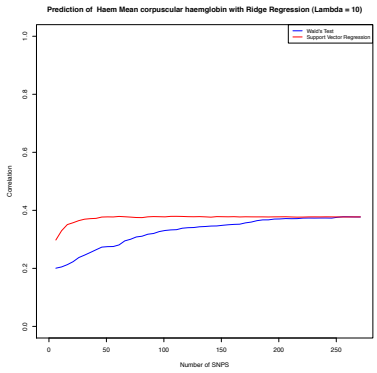
C)



D)



E)



F)

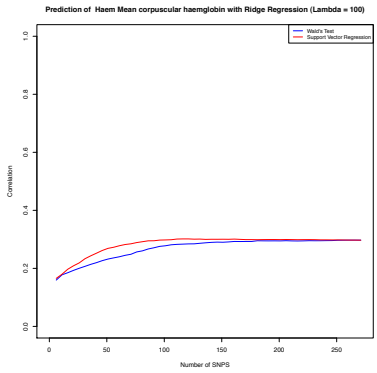
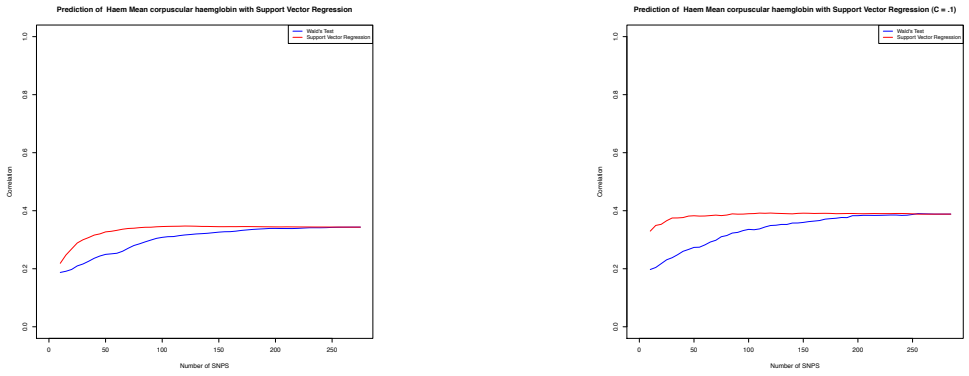


Figure B.2 The prediction of mean cellular hemoglobin using ridge regression including pedigree information with $\lambda = 0$ (A), $\lambda = 0.1$ (B), $\lambda = 1$ (C), $\lambda = 5$ (D), $\lambda = 10$ (E) and $\lambda = 100$ (F) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

Table B.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Mean Cellular Hemoglobin using Support Vector Regression with Pedigree Included at Various Values of *C*

C	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.34 (.02)	265	0.35 (.02)	115	0.17 (.03)	0.18 (.02)	0.18 (.02)
.1	0.39 (.02)	250	0.39 (.02)	115	0.25 (.03)	0.26 (.02)	0.25 (.03)
.01	0.39 (.03)	260	0.39 (.03)	250	0.25 (.03)	0.26 (.02)	0.26 (.03)
.001	0.38 (.02)	260	0.38 (.02)	115	0.24 (.03)	0.25 (.02)	0.25 (.02)
.0001	0.34 (.02)	270	0.34 (.02)	120	0.17 (.03)	0.17 (.02)	0.17 (.02)
.00001	0.25 (.03)	265	0.25 (.03)	270	0.13 (.02)	0.13 (.02)	0.13 (.02)

A) **B)**



C) **D)**

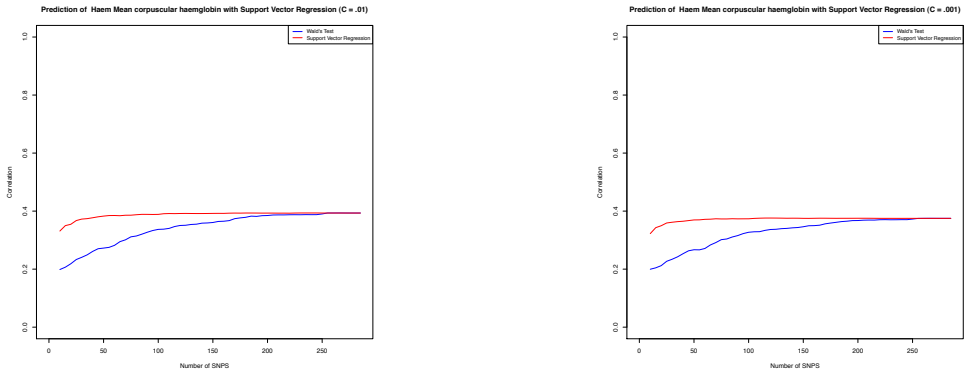


Figure B.3 The prediction of mean cellular hemoglobin using support vector regression including pedigree information with $C = 0$ (A), $C = .1$ (B), $C = .01$ (C) and $C = .001$ (D) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue).

A) **B)**

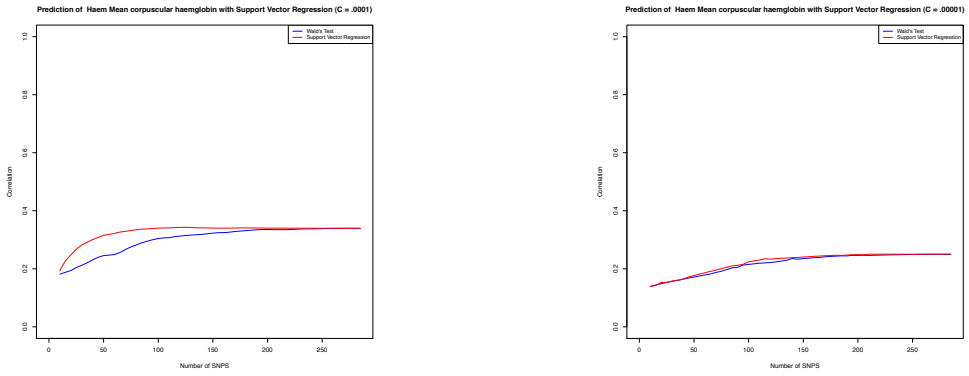


Figure B.4 The prediction of mean cellular hemoglobin using support vector regression including pedigree information with C = .0001 (A), C = .00001 (B) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

APPENDIX C

PERCENTAGE OF $CD8^+$ CELLS ANALYSIS

This appendix contains information of the various parameters used in each of the prediction methods for percentage of $CD8^+$ cells with the figures corresponding to the preceding tables.

Table C.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Ridge Regression with no Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.40 (0.04)	100	0.60 (0.02)	110	0.28 (0.04)	0.34 (0.05)	0.41 (0.04)
.1	0.19 (0.10)	30	0.50 (0.07)	45	0.28 (0.05)	0.33 (0.05)	0.42 (0.04)
1	0.41 (0.04)	230	0.61 (0.02)	110	0.29 (0.04)	0.33 (0.05)	0.43 (0.04)
5	0.46 (0.04)	230	0.63 (0.02)	135	0.29 (0.04)	0.33 (0.05)	0.42 (0.04)
10	0.47 (0.04)	265	0.62 (0.02)	115	0.29 (0.04)	0.33 (0.05)	0.42 (0.04)
100	0.42 (0.06)	320	0.56 (0.03)	335	0.28 (0.04)	0.33 (0.05)	0.42 (0.04)

Table C.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Ridge Regression with Pedigree Included at Various Values of λ

λ	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.40 (0.04)	100	0.60 (0.02)	110	0.29 (0.04)	0.34 (0.05)	0.43 (0.04)
1	0.41 (0.06)	95	0.61 (0.02)	110	0.29 (0.04)	0.34 (0.05)	0.43 (0.04)
5	0.46 (0.04)	230	0.64 (0.02)	135	0.29 (0.04)	0.33 (0.05)	0.41 (0.04)
10	0.47 (0.04)	265	0.62 (0.02)	135	0.29 (0.04)	0.34 (0.05)	0.42 (0.04)
100	0.42 (0.05)	320	0.56 (0.03)	335	0.28 (0.05)	0.35 (0.03)	0.41 (0.04)

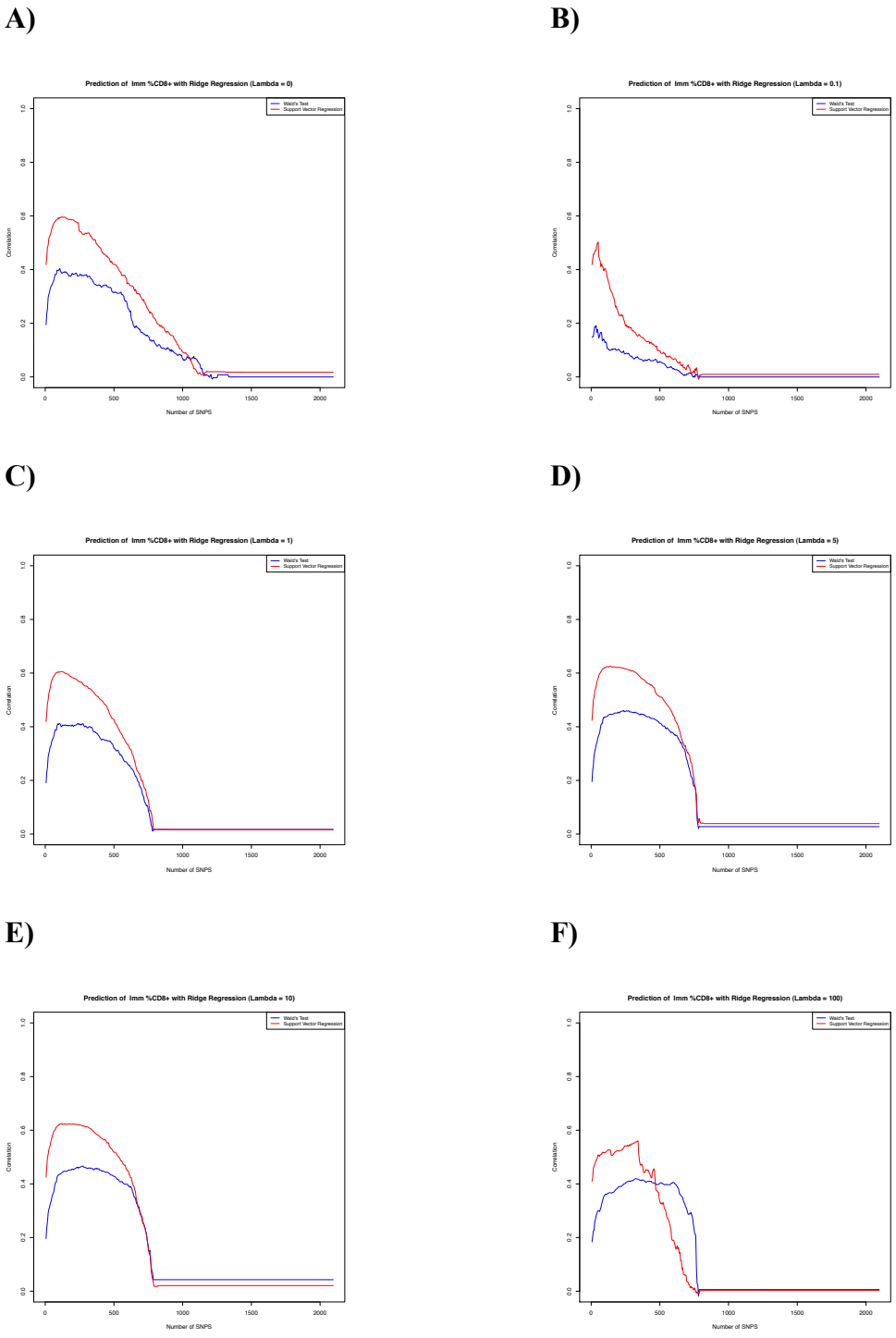


Figure C.1 The prediction of percentage of $CD8^+$ cells using ridge regression without including pedigree with $\lambda = 0$ (A), $\lambda = 0.1$ (B), $\lambda = 1$ (C), $\lambda = 5$ (D), $\lambda = 10$ (E) and $\lambda = 100$ (F) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

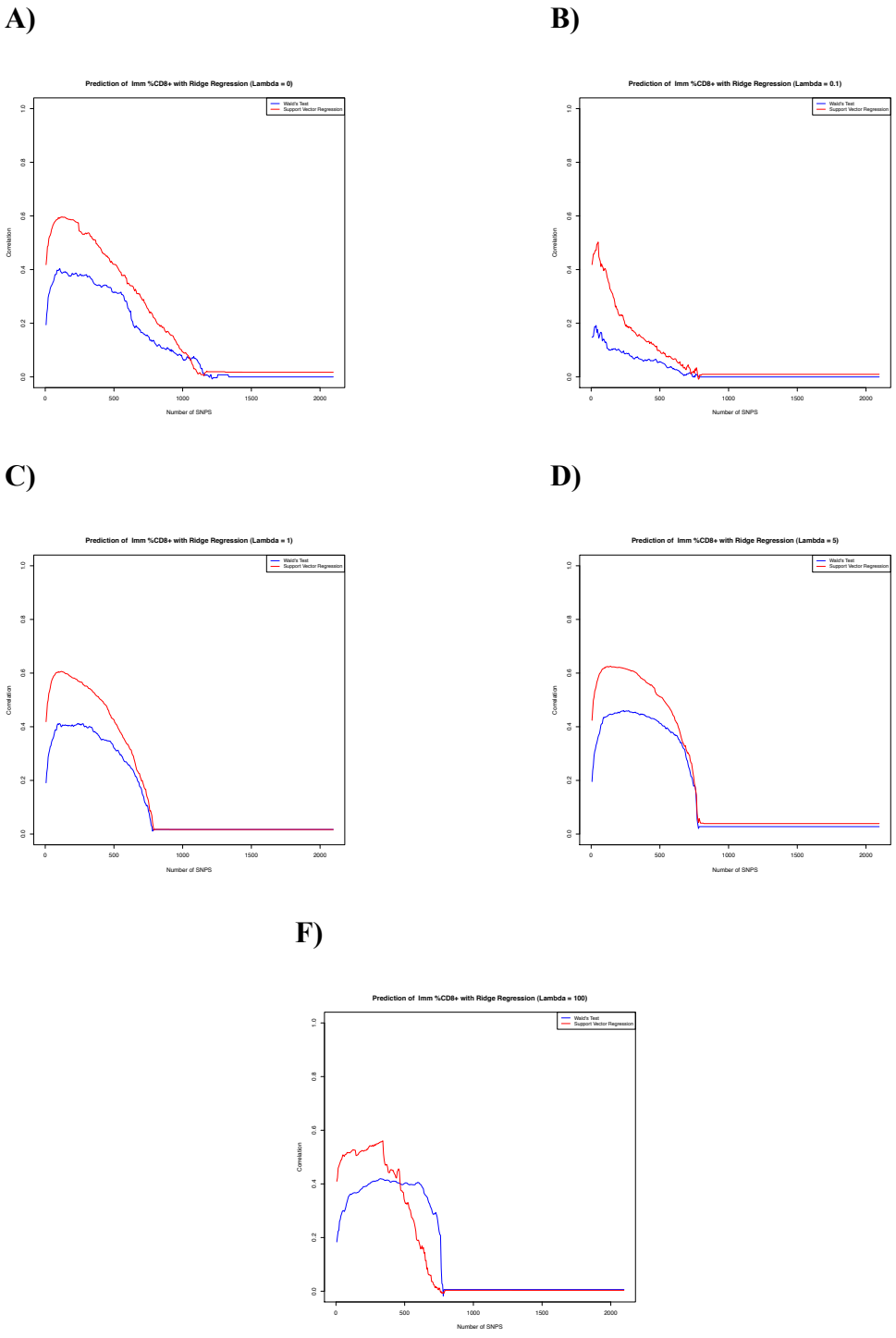
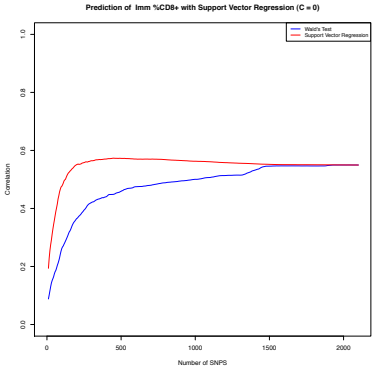


Figure C.2 The prediction of percentage of $CD8^+$ cells using ridge regression including pedigree with $\lambda = 0$ (A), $\lambda = 1$ (B), $\lambda = 5$ (C), $\lambda = 10$ (D) and $\lambda = 100$ (E) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

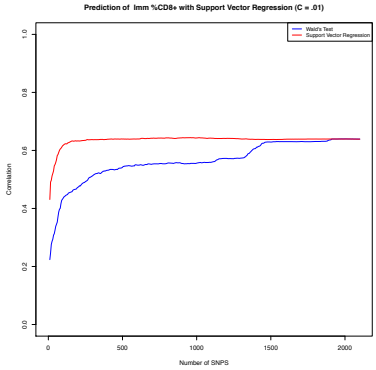
Table C.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for percentage of $CD8^+$ cells using Support Vector Regression with Pedigree Included at Various Values of *C*

C	Wald Test	#	SVR	#	5 Clusters	10 Clusters	20 Clusters
0	0.55 (.03)	2030	0.57 (.03)	445	0.07 (.02)	0.09 (.03)	0.12 (.03)
.01	0.64 (.02)	2000	0.64 (.01)	955	0.26 (.04)	0.32 (.04)	0.40 (.04)
.001	0.63 (.02)	2090	0.63 (.02)	2090	0.14 (.04)	0.21 (.04)	0.30 (.05)
.0001	0.55 (.03)	2030	0.57 (.03)	730	0.07 (.02)	0.08 (.02)	0.11 (.02)

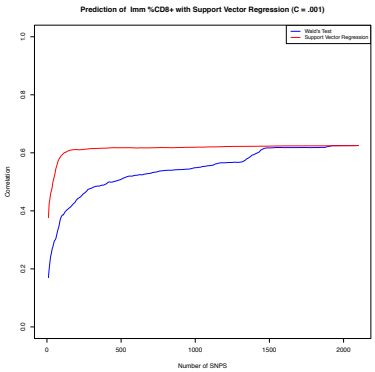
A)



B)



C)



D)

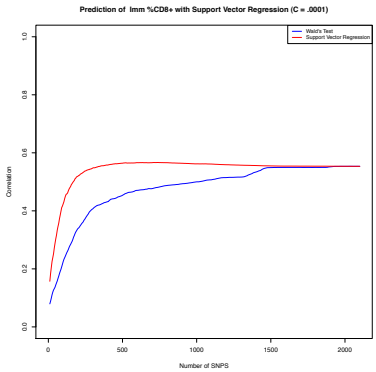


Figure C.3 The prediction of percentage of $CD8^+$ cells using support vector regression including pedigree with $C = 0$ (A), $C = .01$ (B), $C = .001$ (C) and $C = .0001$ (D) shows SVR-ranked SNPs (red) outperforming Wald Test-ranked SNPs (blue) by achieving a maximum correlation, between actual and predicted phenotype, with fewer SNPs.

APPENDIX D

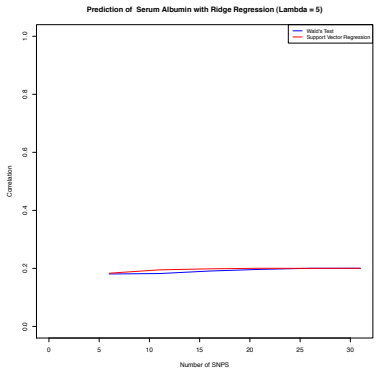
ANALYSIS OF OTHER PHENOTYPES

Confirmation of the SVR-method improvement was performed on other phenotypes that were available [6]. The other phenotypes were showcased that very few SNPs or no SNPs within the Bonferroni correction, a minimum of the 25 SNPs with the highest p-value was chosen, no visible improvement was observed.

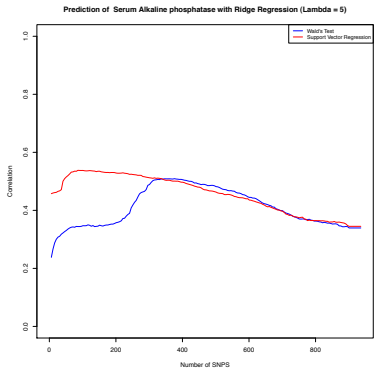
Table D.1 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Biochemical Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$

Phenotype	Max SNPs	Wald Test	#	SVR	#
Albumin	25 (1)	0.20 (0.01)	25	0.20 (0.01)	15
ALP	716 (108)	0.51 (0.02)	370	0.54 (0.02)	80
ALT	25 (0)	0.31 (0.03)	0	0.29 (0.04)	0
AST	25 (0)	0.24 (0.02)	0	0.23 (0.03)	0
Calcium	25 (0)	0.23 (0.02)	5	0.24 (0.01)	0
Chloride	25 (0)	0.26 (0.01)	15	0.25 (0.02)	15
Creatinine	25 (0)	0.21 (0.03)	0	0.20 (0.02)	0
Glucose	25 (0)	0.11 (0.04)	20	0.11 (0.04)	20
HDL	312 (81)	0.42 (0.02)	80	0.44 (0.02)	80
LDL	40 (14)	0.30 (0.03)	50	0.31 (0.03)	35
Phosphorous	25 (0)	0.15 (0.04)	20	0.15 (0.04)	5
Potassium	25 (0)	0.09 (0.15)	5	0.12 (0.10)	0
Sodium	25 (0)	0.21 (0.03)	5	0.21 (0.03)	0
Tot.Cholesterol	67 (23)	0.36 (0.02)	95	0.36 (0.02)	30
Tot.Protein	25 (0)	0.12 (0.04)	20	0.12 (0.04)	20
Triglycerides	25 (0)	0.14 (0.05)	20	0.14 (0.05)	15
Urea	264 (72)	0.27 (0.03)	200	0.28 (0.04)	70

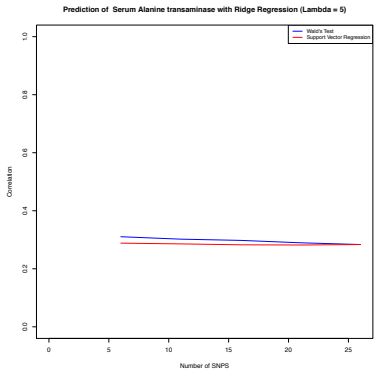
A)



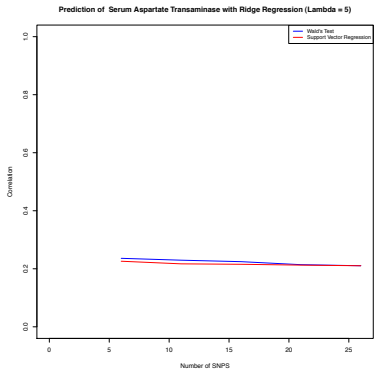
B)



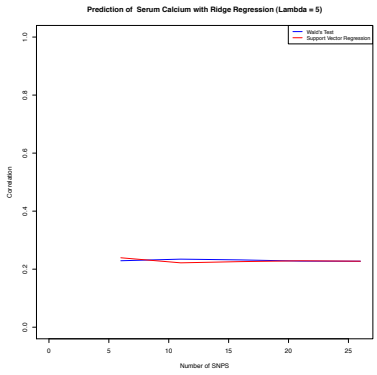
C)



D)



E)



F)

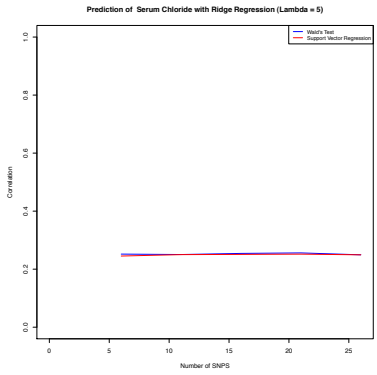


Figure D.1 The prediction of Albumin (A), ALP (B), ALT (C), AST (D), Calcium (E), and Chloride (F) using ridge regression including pedigree with $\lambda = 5$.

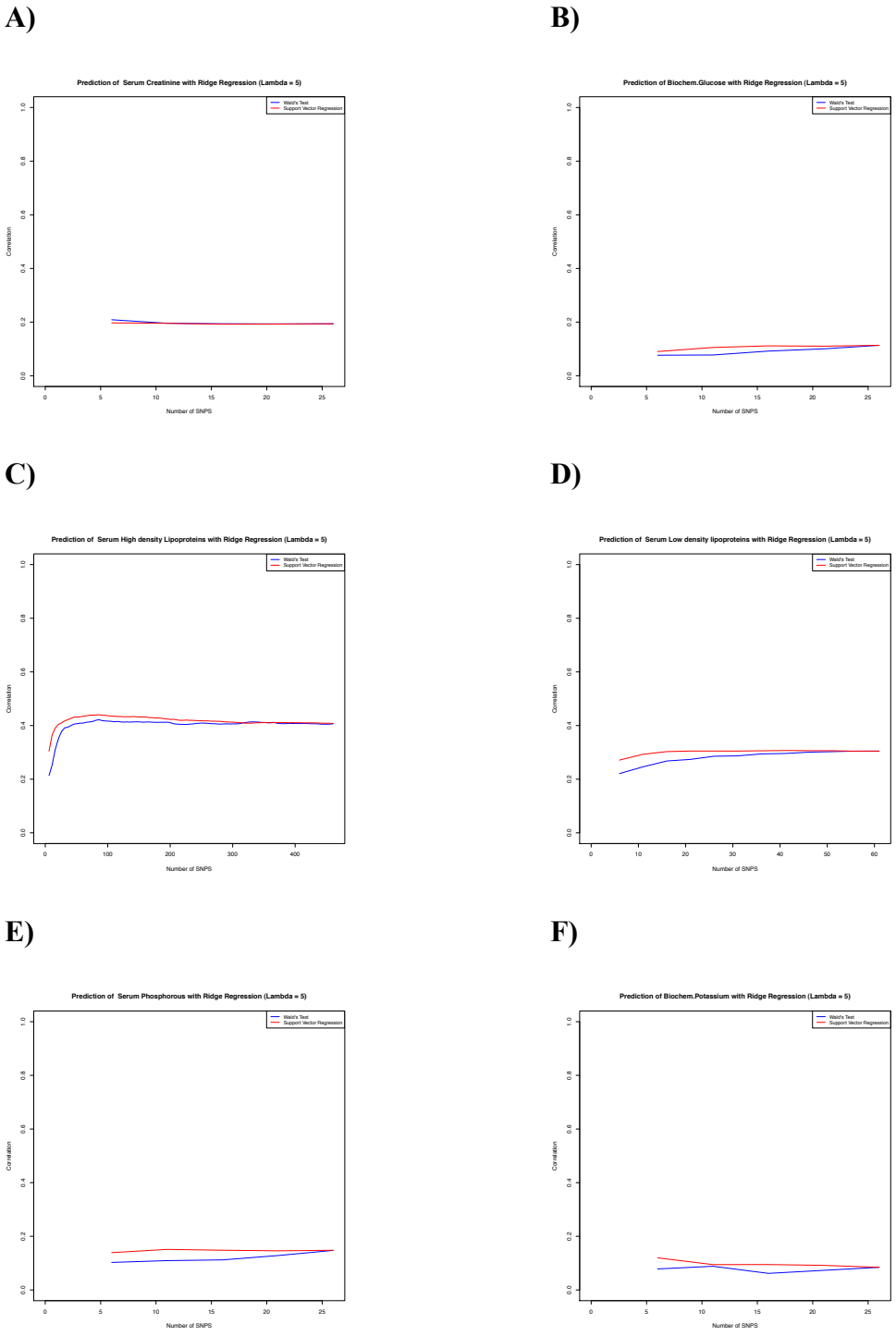
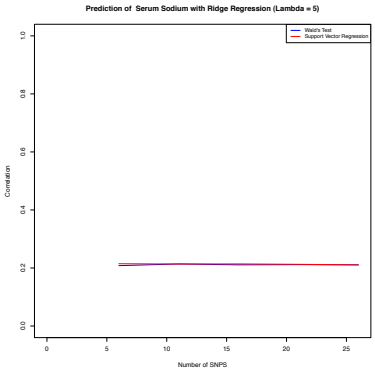
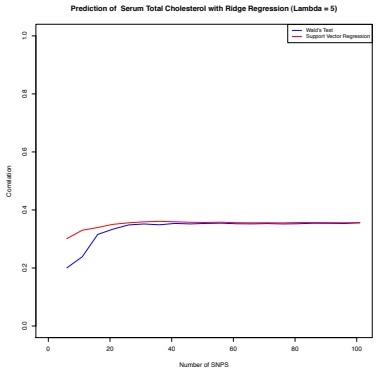


Figure D.2 The prediction of Creatinine (A), Glucose (B), HDL (C), LDL (D), Phosphorus (E) and Potassium (F) using ridge regression including pedigree with $\lambda = 5$.

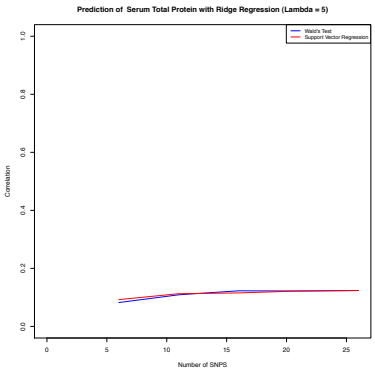
A)



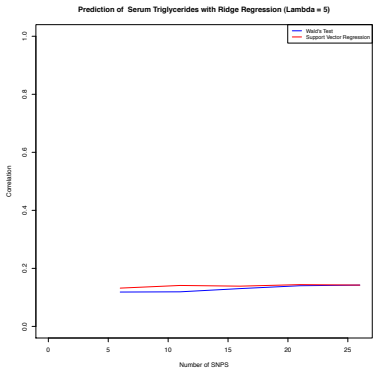
B)



C)



D)



F)

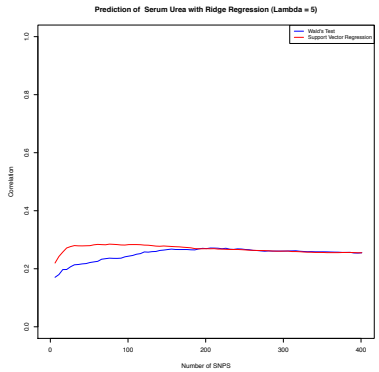


Figure D.3 The prediction of Sodium (A), Tot.Cholesterol (B), Tot.Protein (C), Triglycerides (D) and Urea (E) using ridge regression including pedigree with $\lambda = 5$.

Table D.2 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Immunological Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$

Phenotype	Max SNPs	Wald Test	#	SVR	#
B220Median	103 (35)	0.29 (0.04)	165	0.29 (0.03)	85
CD4XGeoMean	474 (128)	0.58 (0.02)	190	0.59 (0.02)	115
CD4YGeoMean	25 (0)	0.20 (0.03)	20	0.20 (0.03)	5
CD4inCD3XGeoMean	530 (103)	0.58 (0.01)	165	0.60 (0.02)	95
CD4inCD3YGeoMean	25 (0)	0.20 (0.04)	20	0.20 (0.04)	15
CD8XGeoMean	25 (0)	0.18 (0.03)	20	0.18 (0.03)	15
CD8YGeoMean	32 (11)	0.32 (0.02)	45	0.32 (0.02)	30
CD8inCD3XGeoMean	27 (5)	0.18 (0.06)	5	0.18 (0.05)	15
CD8inCD3YGeoMean	32 (13)	0.31 (0.02)	40	0.31 (0.02)	50
PctB220	327 (44)	0.43 (0.01)	310	0.44 (0.02)	170
PctCD3	260 (114)	0.40 (0.02)	215	0.42 (0.01)	115
PctCD4	179 (56)	0.34 (0.02)	220	0.36 (0.02)	80
PctCD4inCD3	1284 (223)	0.42 (0.04)	110	0.59 (0.02)	75
PctCD8inCD3	1563 (206)	0.43 (0.02)	155	0.63 (0.02)	90

A) **B)**

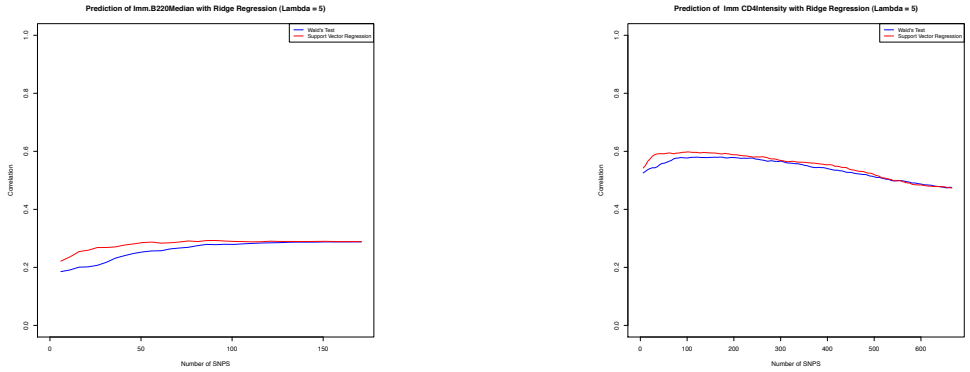


Figure D.4 The prediction of B220Median (A) and CD4XGeoMean (B) using ridge regression including pedigree with $\lambda = 5$.

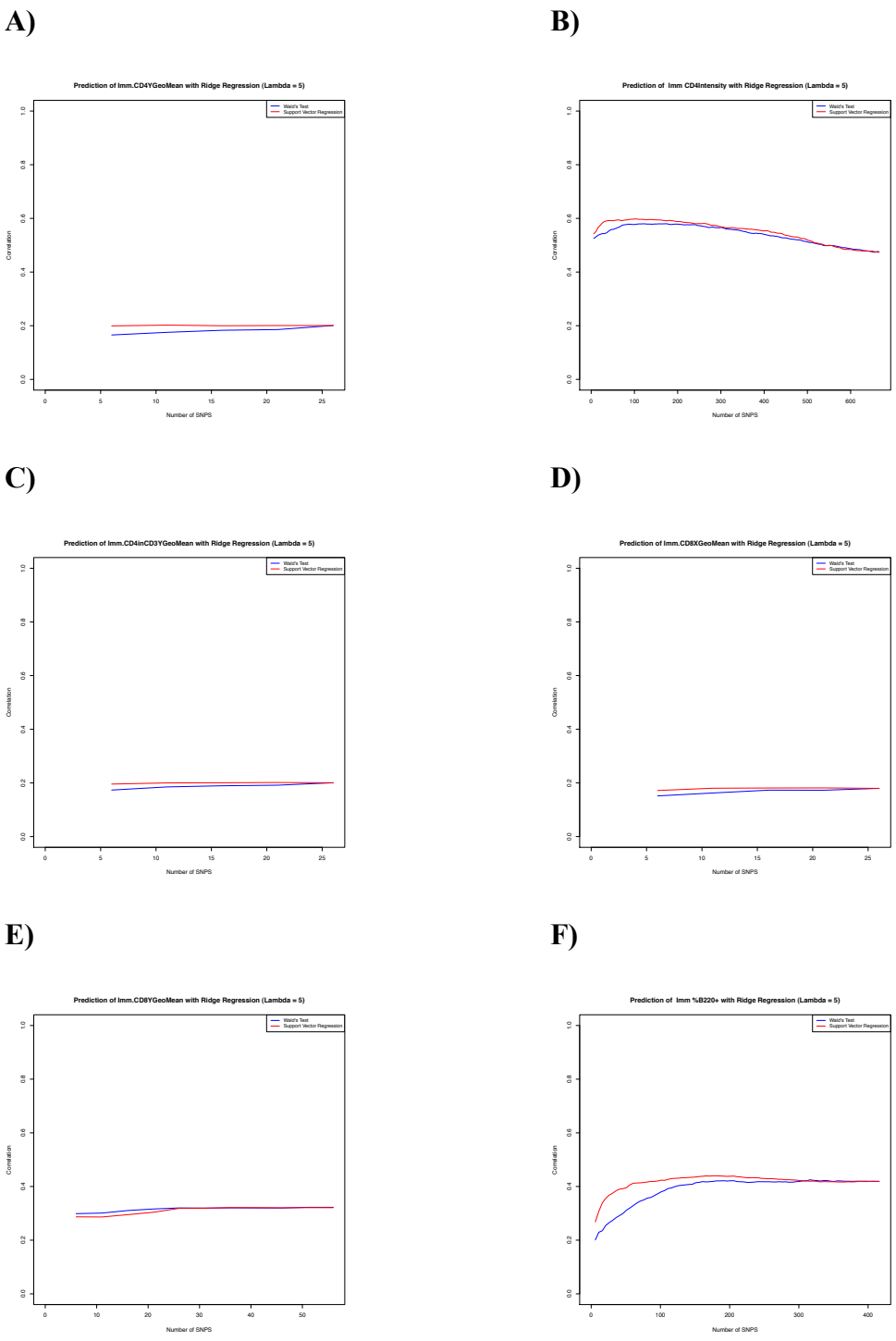
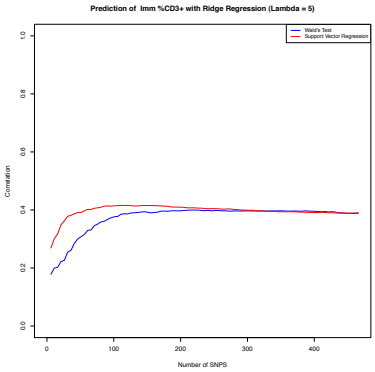
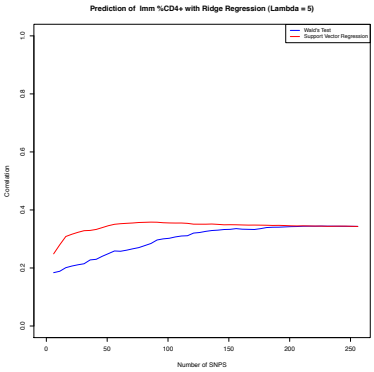


Figure D.5 The prediction of CD4YGeoMean (A), CD4inCD3XGeoMean (B), CD4inCD3YGeoMean (C), CD8XGeoMean (D), CD8YGeoMean (E) and PctB220 (F) using ridge regression including pedigree with $\lambda = 5$.

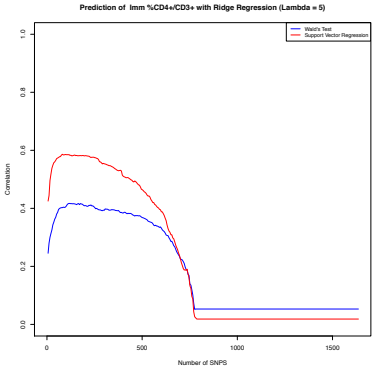
A)



B)



C)



D)

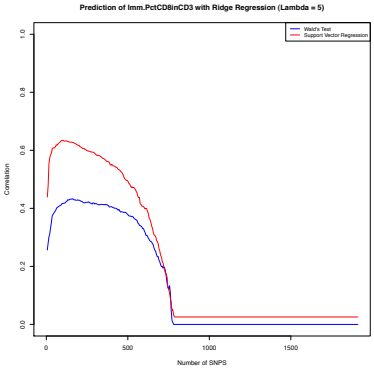


Figure D.6 The prediction of PctCD3 (A), PctCD4 (B), PctCD4inCD3 (C) and PctCD8inCD3 (D) using ridge regression including pedigree with $\lambda = 5$.

Table D.3 Maximum Correlation (and Standard Deviation) between Actual and Predicted Values for Hematological Phenotypes using Ridge Regression with Pedigree Included at $\lambda = 5$

Phenotype	Max SNPs	Wald Test	#	SVR	#
ALYabs	25 (0)	0.09 (0.03)	15	0.09 (0.03)	5
BASabs	25 (0)	0.09 (0.04)	20	0.09 (0.04)	20
HCT	25 (0)	0.10 (0.03)	15	0.11 (0.04)	0
HGB	25 (1)	0.15 (0.03)	20	0.15 (0.02)	5
LICabs	25 (0)	0.11 (0.02)	0	0.10 (0.02)	15
LYMabs	86 (47)	0.33 (0.02)	85	0.33 (0.02)	25
MCHC	147 (49)	0.38 (0.03)	155	0.39 (0.02)	45
MCV	602 (121)	0.38 (0.03)	275	0.46 (0.02)	145
MONabs	25 (0)	0.21 (0.02)	20	0.21 (0.02)	20
MPV	54 (14)	0.35 (0.02)	45	0.35 (0.02)	10
NEUabs	40 (16)	0.21 (0.04)	55	0.22 (0.04)	10
PCT	25 (0)	0.13 (0.02)	20	0.13 (0.02)	15
PLT	25 (0)	0.18 (0.02)	20	0.19 (0.02)	10
RBC	26 (3)	0.15 (0.02)	15	0.14 (0.02)	20
RDW	172 (46)	0.43 (0.02)	195	0.45 (0.02)	70
WBC	74 (39)	0.30 (0.02)	90	0.30 (0.02)	60

A) **B)**

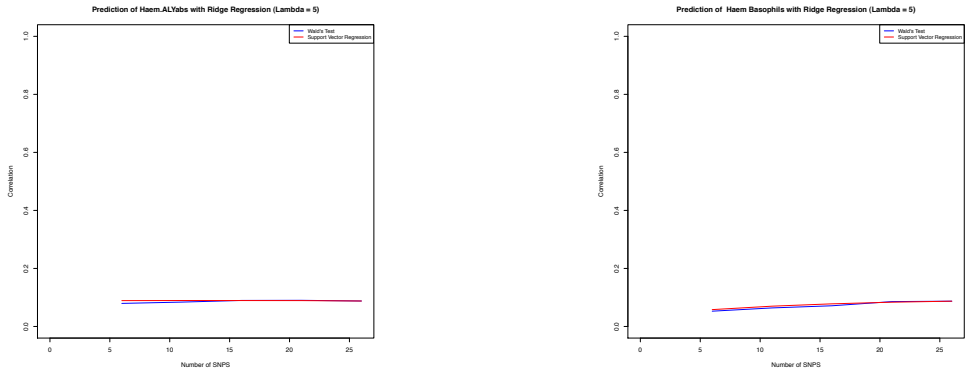


Figure D.7 The prediction of ALYabs (A) and BASabs (B) using ridge regression including pedigree with $\lambda = 5$.

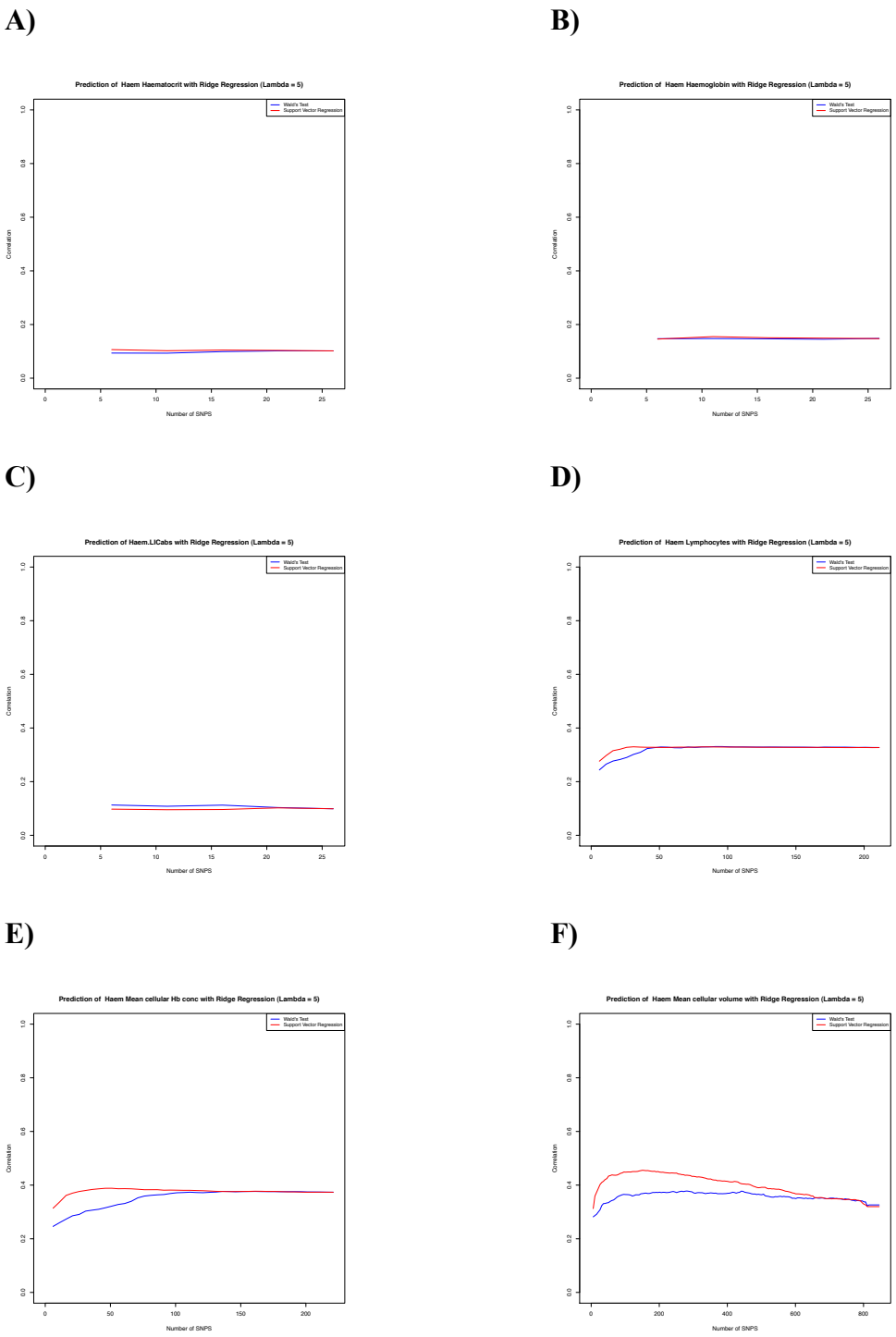


Figure D.8 The prediction of HCT (A), HGB (B), LICabs (C), LYMabs (D), MCHC (E) and MCV (F) using ridge regression including pedigree with $\lambda = 5$.

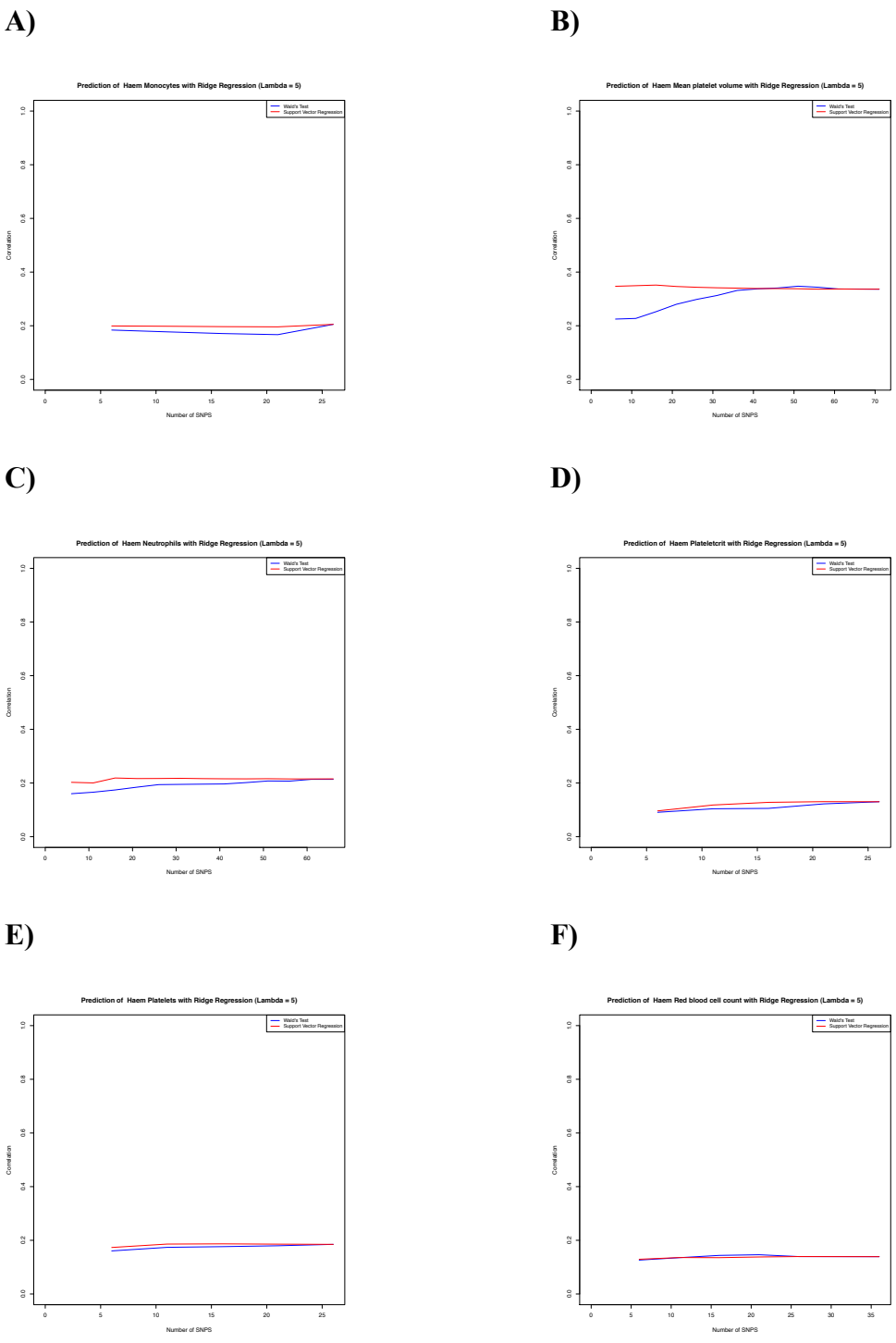


Figure D.9 The prediction of MONabs (A), MPV (B), NEUabs (C), PCT (D), PLT (E) and RBC (F) using ridge regression including pedigree with $\lambda = 5$.

A) **B)**

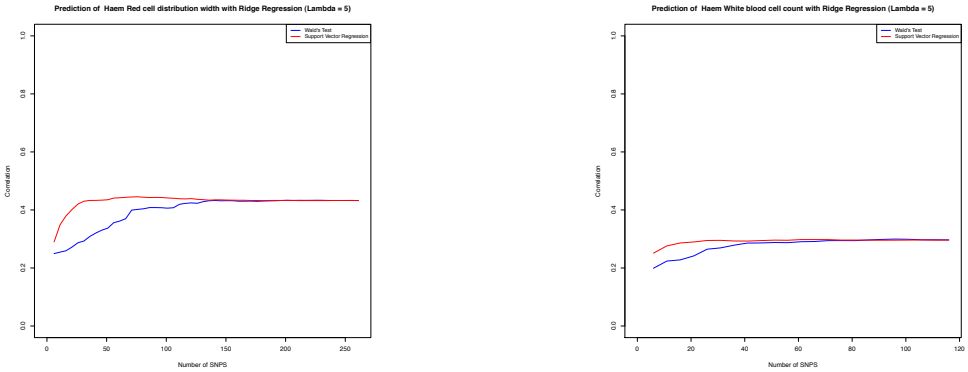


Figure D.10 The prediction of RDW (A) and WBC (B) using ridge regression including pedigree with $\lambda = 5$.

APPENDIX E

RIDGE REGRESSION RESULTS

The following table contains the prediction values of the first 25 individuals of one trial of the MCH phenotype with 850 SNPs selected for ridge regression. The first 25 individuals illustrate the immense difference between predicted and actual phenotype that causes the correlation to deteriorate past 800 SNPs.

Table E.1 Actual and Predicted Values of Phenotypes with Ridge Regression

Actual Phenotype	Predicted Phenotype
14.4	16.32767027
15.7	19.98244374
14.8	-0.361602382
15.9	40.69157345
17.1	29.00447911
15.9	11.8423628
15.5	18.50291378
15.1	-1.632734752
15.3	5.666539956
15.7	40.56899092
16.1	20.19513194
14.8	25.08874955
16.6	29.6820226
15.6	10.98136206
15.7	20.17615257
15.5	-18.05698297
15.6	13.21237284
15.2	43.27339969
15.9	6.44208568
15.3	-0.491817296
15.9	12.16342734
16.2	20.01154506
15.1	37.50043935
14.7	-6.244673477
15.7	18.25904409

REFERENCES

1. Stromberg,U., Bjork,J., Vineis,P., Broberg,K. and Zeggini,E. (2009) Ranking of genome-wide association scan signals by different measures. *International Journal of Epidemiology*, **38**, 1364-1373.

2. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H. (2011) Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Research,* **1-8**. doi:10.1093|nar|gkr064

3. Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscer PM. (2008) Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. *PLoS Genetics*, **4**(10): e1000231. doi:10.1371/journal.pgen.1000231.

4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**.

5. Malo N, Libiger O and Schork NJ. (2008) Accommodating Linkage Disequilibrium in Genetic-Association Analyses via Ridge Regression. *The American Journal of Human Genetics*, **82**, 375-385.

6. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, et al. (2006) Genetic and Environmental Effects on Complex Traits in Mice. *Genetics,* **174**: 959-984.

7. Joachims,T. (1999) Making large-scale svm learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA.

8. Smola AJ, Schölkopf B. (2003) A Tutorial on Support Vector Regression.

9. K. Crammer and Y. Singer. (2001) On the Algorithmic Implementation of Multi-class SVMs, *Journal of Machine Learning Research*.